

格致方法·定量研究系列 吴晓刚 主编

指数随机图模型导论

[美] 詹宁·K. 哈瑞斯（Jenine K. Harris）著
杨冠灿 译

 革新研究理念

 丰富研究工具

 最权威、最前沿的定量研究方法指南

51

格致出版社 上海人民出版社



本书介绍了指数随机图模型的基本概念，通过案例解释了为什么要使用指数随机图模型，并向读者展现了如何在研究中运用基本的指数随机图模型进行分析。

指数随机图模型是一种针对社会网络结构进行建模的统计方法。近年来，随着统计软件的不断改进，已经有一些社会科学家开始使用指数随机图模型统计工具进行研究。然而，目前尚缺乏一本精炼的模型使用指南。本书通过公共健康领域的真实案例以及详细指导读者使用 R 统计软件及 statnet 包，填补了这一空白。

主要特点

- 指数随机图模型是一种网络统计推
- 本书采用了公共健康领域的真实案例，并详细讲解了分析流程
- 本书提供了指数随机图建模的完整 R 语言代码，用户可以方便地重复完整的指数随机图建模过程

上架建议：社会研究方法

ISBN 978-7-5432-2654-8



定价：30.00元

您可以通过如下方式联系到我们：
邮箱：hibooks@hibooks.cn


微信


天猫

易文网：www.ewen.co
格致网：www.hibooks.c

格致方法·定量研究系列 吴晓刚 主编

指数随机图模型导论

[美] 詹宁·K. 哈瑞斯（Jenine K. Harris） 著
杨冠灿 译

SAGE Publications, Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

指数随机图模型导论/(美)詹宁·K.哈瑞斯著；杨冠灿译.—上海：格致出版社：上海人民出版社，2016.10

(格致方法·定量研究系列)

ISBN 978-7-5432-2654-8

Ⅰ.①指… Ⅱ.①詹… ②杨… Ⅲ.①社会关系-指数模型-研究 Ⅳ.①C912.3②F224.0

中国版本图书馆 CIP 数据核字(2016)第 184147 号

责任编辑 张苗凤

格致方法·定量研究系列

指数随机图模型导论

[美]詹宁·K.哈瑞斯 著
杨冠灿 译

出 版	世纪出版股份有限公司 格致出版社 世纪出版集团 上海人民出版社 (200001 上海福建中路193号 www.ewen.co)	印 刷	浙江临安曙光印务有限公司
	编辑部热线 021-63914988 市场部热线 021-63914081 www.hibooks.cn	开 本	920×1168 1/32
		印 张	6
		字 数	120,000
发 行	上海世纪出版股份有限公司发行中心	版 次	2016年10月第1版
		印 次	2016年10月第1次印刷

ISBN 978-7-5432-2654-8/C·153 定价：30.00元



出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于2011年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于2012年以单行本的形式再次出版发行，共37本。我们衷心感谢广大读者的支持和建议。

随着与SAGE出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

2003年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复，但各有侧重。"社会科学里的统计学"从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。"社会科学中的定量分析"则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的"定量社会科学研究"丛书闻名遐迩，每本书都写得通俗易懂，与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在"格致方法・定量研究系列"丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现"化整为零、循序渐进、中文教学、学以致用"的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法 · 定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

自20世纪初乔治·齐美尔(Georg Simmel)首次论述社会网络相关问题以来(Simmel & Wolff, 1950),社会科学家对于个体之间、组织之间以及其他实体之间相互关联的网络问题一直保持高度的关注(参见例如Fienberg, 2012)。20世纪30年代,心理医生雅各布·莫雷诺(Jacob Moreno, 1934)的工作为社会网络研究奠定了基础,并将此领域命名为"社会计量学"(sociometry)。在莫雷诺的诸多重要成果中,核心成果便是发明了社群图(sociogram)方法,通过将个体图形化表示为节点,个体之间联系图形化表示为连线的形式,社群图方法就能够用来解释社会结构问题。

在社会网络分析发展的历程中,社群图方法被证明是十分重要的,原因之一是社群图方法将图论的基础理论引入到了社会网络分析中来。图论是一个专门处理由节点(点)以及相连的边(连线)所组成的数学分支,其中,网络图既可以是有向的,即网络中的边通常由从一个节点到另一个节点的箭头所表示,从而展现节点之间潜在的非对称联系;网络图也可以是无向的,直接用线段来表示网络中的边。大多数研

究社会网络的传统方法都是来源于图论的，社会科学中的定量研究方法应用系列丛书(QASS)中，有一本较早的著作，是由诺克和杨(Knoke & Yang, 2008)撰写的《社会网络分析》，该书就主要是采用这种(传统)方法。

传统的网络分析方法主要是描述性的，并不采用具有统计学意义上的随机变量模型构建方法。明确提出以网络结构为中心建立概率模型的思想可以追溯到20世纪中叶，即吉尔伯特、艾多斯以及瑞尼(Gilbert, 1959; Erdos & Renyi, 1959)解释了网络结构中最为基础的零模型(null model)。在零模型中，所有的节点对都是以同等的概率建立连线，无论是在有向网络还是无向网络中，简单图模型都是被最广泛采用的模型。

20年之后，霍兰德和莱因哈特(Holland & Leindardt, 1981)引入了一种针对有向图的Gilbert-Erdos-Renyi零模型的变种。其中，关系形成(tie formation)的概率受到个体的群集性(gregariousness，个体对外与他人建立联系的属性)以及受欢迎程度(popularity，他人与该个体建立联系的属性)的影响。在此之后不久，1981年，芬博格和沃瑟曼(Fienberg & Wasserman, 1981)将霍兰德和莱因哈特的 p_1 模型改造为对数线性模型，对数线性模型是一种为统计学家和社会科学家所熟知的模型，这样一来，学者们就可以方便对模型的参数进行最大似然估计了。此外，芬博格和沃瑟曼还对 p_1 模型进行了扩展，将网络的“互惠性”(reciprocity)特征纳入到模型中来，并以“互惠性”特征作为网络连线概率增强的机制——例如，在一个朋友网络中，如果A选择B，那么，B选择A的概率就会提升。

正如詹宁·哈瑞斯(Jenine Harris)在本书中所解释的，吉尔伯特等人的零模型、霍兰德和莱因哈特的 p_1 模型，以及芬博格和沃瑟曼(1981)的扩展模型都是指数随机图模型(exponential random graph models，ERGMs)家族的成员。过去30年里，指数随机图模型的研究取得了长足的进展，而且已经成为了目前社会网络分析中最重要的统计工具。在这个进程中，指数随机图模型不断彰显着自己在展现社会网络结构特征分析方面的洞察力，例如对聚类或“聚簇”的分析。

近年来，面对大数据分析所带来的挑战与激励，计算机科学家和统计物理学家，与统计学家、社会科学家并肩作战，对社会网络分析的发展起到了直接推动作用。源于社会生活中的大型网络数据尤为庞大与复杂，如Facebook的数据，这也促使研究人员必须不断研究更为复杂的网络模型，不断改进统计软件的计算能力，以确保研究的模型能够适应大数据的环境。哈瑞斯在其书中介绍了由statnet团队所研发的最先进的网络分析软件(Handcock et al.，2003)，该软件是针对R的统计计算环境而开发的(R Core Team，2013)，是一款广泛使用的、免费且开源的统计分析平台。

本书介绍了如何建立指数随机图模型，并解释了如何在实践中使用该模型，詹宁·哈瑞斯的工作对于采用社会网络分析的社会学家而言十分重要。我希望她的这本著作将会有较广泛的读者群，同时，期待该书能够对社会科学中社会网络分析质量的提升产生实质性的影响。

约翰·福克斯(John Fox)

参考文献

Erdös, P. and Rènyi, A.(1959). On random graphs, I. *Publicationes Mathematicae*, 6:290—297.

Fienberg, S.E.(2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21:825—839.

Fienberg, S.E. and Wasserman, S.S.(1981). Categorical data analysis of single sociometric relations. In Leinhardt, S., editor, *Sociological Methodology 1981*, pages 156—192. Jossey-Bass, San Francisco.

Gilbert, E.N.(1959). Random graphs. *The Annals of Mathematical Statistics*, 30:1141—1144.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M.(2003). *statnet: Software tools for the Statistical Modeling of Network Data*. Seattle, WA.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs(with discussion). *Journal of the American Statistical Association*, 76:33—65.

Knoke, D. and Yang, S.(2008). *Social Network Analysis*. Thousand Oaks CA, second edition.

Moreno, J.(1934). *Who Shall Survive*? Nervous and Mental Disease Publishing Company, Wasington DC.

R Core Team(2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Simmel, G. and Wolff, K.H.(1950). *The Socology of Georg Simmel*. The Free Press, New York.

目 录

第1章

网络分析方法的希望与挑战

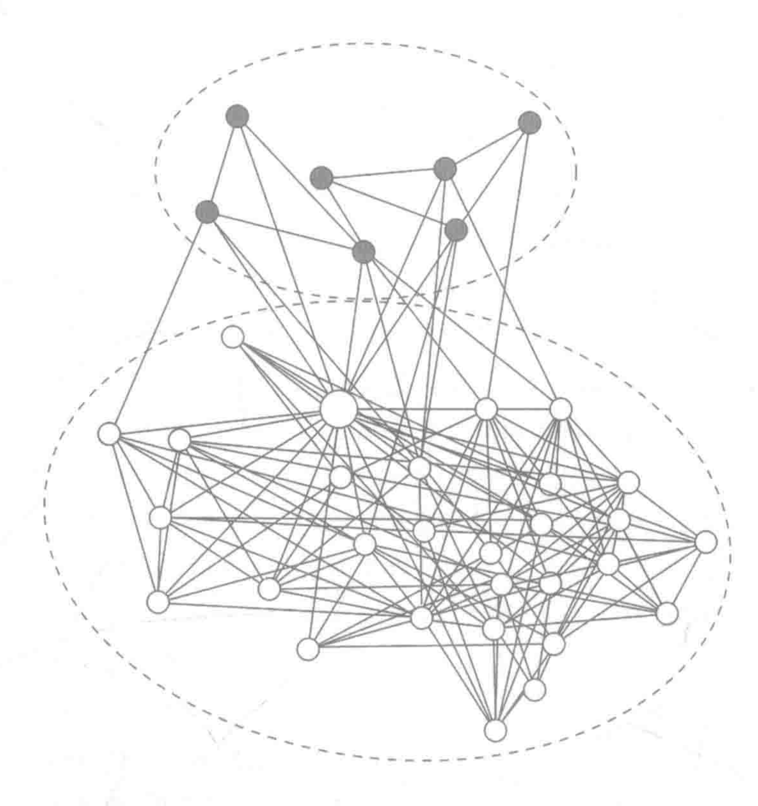
近30年来，经验社会研究被抽样调查方法所主导。但正如人们经常指出的那样，通过对个体进行随机抽样的调查方法成为了一台社会学研究的绞肉机，它把个体从其所在的社会情境中剥离开来，使得研究过程中任何个体之间不存在任何关系。这就有点像一个生物学家让他们的实验动物首先经过一台绞肉机的处理，然后，以百计的细胞为观测对象，通过显微镜对（实验动物）进行观察。这样，解剖学和生理学用不着了，结构与功能消失了，唯一剩下的只有细胞生物学……如果我们的目标是去理解人类的行为而不是简单地记录这些行为，那么，我们希望知道诸如主要群体，邻居关系、组织结构、社会圈子、社区关系等问题；还包括交流、沟通、角色期望以及社会控制等问题。

——巴顿(Barton)，1968，

援引自弗里曼(Freeman，2004:1)

2010年，南亚的一个内陆小国不丹采用了一种用于测量国家成就的新措施——国民幸福总值(GNH)指数，与之前通用的评价国家生产力的经济指标不同，GNH指数侧重于观

察国民的福祉，“意在通过改进尚未获得幸福感（not-yet-happy）人群的生存状况，引导人民和国家走向幸福”（Ura，Alkire，Zangmo & Wangdi，2012）。在科学文献中，GNH也被称为主观幸福感（subjective well-being，SWB），通过总结数十年对于幸福问题的研究经验，GNH指数初步形成为一个包含9大领域、13类指标以及124项变量的综合指标体系。GNH指数所涉变量的范围覆盖了整个生态模型，从个体层次人口统计指标，如年龄与教育程度，到中观层次对家庭紧密度、社区融入度的测量，甚至包括对自然和经济环境的认知等。然而，在GNH指数所包含的若干测量指标中却遗漏了对特定社会关系的测量。之前的研究发现：相邻个体之间的关系数量、关系属性以及关系之间的同质性（其中，同质性是指人们更愿意与其相似的人群进行联系的行为准则）对于个体知识、态度以及交往活动都起到了重要的作用（McPherson，Smith-Lovin & Cook，2001），这种作用也适用于针对幸福感问题的研究（Burt，1987；Myers & Diener，1995）。例如，最近对于幸福感问题的研究就发现，至少在我们的在线网络中，感觉幸福的人总是愿意与那些同样感觉到幸福的人交朋友（Bliss，Kloumann，Harris，Danforth & Dodds，2012；Bollen，Goncalves，Ruan & Mao，2011）。图1.1显示了一个在线交友网络，其中的朋友关系以及他们是否感觉到幸福的状态被标记出来，感觉到幸福的人们被标注为具有较高的主观幸福感，而感觉不幸福的人则具有较低的主观幸福感。



资料来源：改编自 Bollen，Goncalves，Ruan & Mao(2011)。

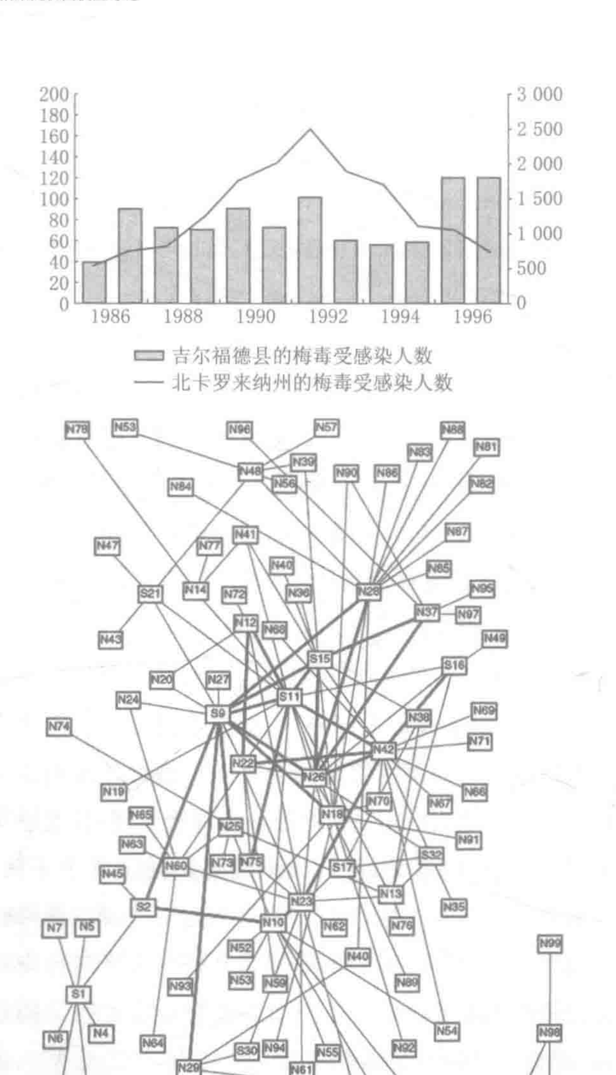
图 1.1 在线交友关系网络以及人们对幸福的感受

因此，想要理解幸福感并识别出那些尚未获得幸福感的人群，就需要从个体社会网络的视角对幸福产生的模式进行认知。然而，促使人们相互联系的纽带不仅仅只涉及幸福感本身，还涉及人们对诸如吸烟行为、就业前景、寿命长短、药物使用、体育运动、书籍选择、健康观念等事物的认知与态度，以及对行政权力、科学发展、疾病传播以及难以计数的与人类生存相关问题的理解（Ennett & Bauman，1993；

Granovetter，1983；Hall & Valente，2007；Harris，Luke，Zuckerman & Shelton，2009；Harris，Carothers，Wald，Shelton & Leischow，2012；Krebs，2000；Luke & Harris，2007；Seeman，Kaplan，Knudsen，Cohen & Guralnik，1987；Valente，2010；Valente & Saba，1998；Valente & Vlahov，2001；Voorhees et al.，2005）。社会科学家对上述行为表现出的特征产生了浓厚的兴趣，然而，传统的社会科学所使用的定量分析方法并没有将关系信息纳入到定量分析范畴中来。相反，传统的定量研究方法是依赖于一项关键的（也是必须的）基础假定条件：即研究中的个体是不相关或彼此相互独立的，基于此种假设条件的研究方法可能会导致分析结果遗失许多重要信息。

例如，1997年，北卡罗来纳州吉尔福德县暴发了梅毒传播事件（参见图1.2，上半部分）。标准的流行病监测技术是通过追踪传染疾病案例数量的变化来对梅毒传播范围进行监测的。通过对流行病数据的进一步观察发现：在受感染人群中，年轻人是最多的，而且大多数人还处于青少年阶段。一旦确认了受感染人群的这项特征，当地卫生部门的性传播疾病项目工作人员就决定采用网络分析方法来对传染病暴发进行跟踪与研究。该方法不仅对受感染者进行了问卷调查，还调查了可能与受感染者有关系的人，因为这些人也可能会参与到传染病传播过程中来，无论他们是否被感染。

通过上述网络分析方法，工作人员发现：在当地的99个年轻人群体之间存在一个复杂的性关系网络，而其中有10名年轻人为受感染者。如图1.2下半部分的网络图所示，99个年轻人由方框来表示，他们之间的性关系用连线表示，较



图上半部分展示的是一个传统的流行病监测方法的分析结论(疾病控制与防治中心,1998);图下半部分显示了网络方法的分析结论。两种方法都旨在理解1997年在北卡罗来纳州吉尔福德县暴发的梅毒事件(Rothenberg et al., 1998)。

图 1.2

粗的线条表示已经查明的梅毒传播途径。从网络图中可以清楚地了解到，大多数的年轻人都具有多个性伙伴，这一特点促使疾病迅速蔓延。正是由于网络存在这样的关系模式，无论他们是否受到感染，这些年轻人都有33%的机会通过性关系接触到网络中的其他受感染者。可见，在识别受传染病威胁人群的范围、危险程度，以及关系模式（促使梅毒在社区的个体间、群体间进行迅速传播）方面，最初的分析方法——传染病暴发初期采用的非关系型的方法是不充分的（Rothenberg et al., 1998）。

在面对诸如在线交友过程中幸福感问题以及监测传染病暴发问题时，非传统方法正在挑战那些在社会科学中曾被广泛接受的、传统的数据采集方法与分析技术（Freeman, 2004）。尽管目前已经存在大量呼声要求将情境方法（例如多层次建模、空间统计和网络分析）整合到从社区心理学到信息通讯等多个领域（Green, 2006; Hirsch, Levine & Miller, 2007; Leischow et al., 2008; Luke, 2005; Luke & Stamatakis, 2012; Shumate & Palazzolo, 2010），但社会科学中的大多数研究仍然依赖于标准的研究方法和工具，而这些方法和工具均以观察对象的独立性假设为前提（Luke & Stamatakis, 2012）。为了符合这种假定条件，一个典型研究通常需要获得一个随机独立样本，而且采样过程必须明确是针对无关联个体进行的独立采样。为了符合这种假定所期望的独立性效果，最理想的样本要求个体之间一定不是邻里关系或性伴侣关系，最好也不要共同出入一个教堂，甚至要求最好不要在同一家餐厅用餐等。和非参数统计方法一样，大量的广义线性模型家族也依赖于这种假定。

当然，在理解人类身心健康、社会正义、经济、政治以及人类存在的许多方面，在这些被广泛采用的传统方法也是有益的。

然而，正如巴顿(Barton)明确指出的，独立性假定要求将个体信息从其所依附的情境去除或“剥离”开来，而这些情景因素又被证明对于理解行为和效果具有十分重要的意义。传统的标准方法的最大特征在于，对个体信息进行“去情境化”(de-contextualization)处理，这一特征极大地限制了待检验假设的选择范围及其所蕴含背景知识的选择范围；同样地，这种将个体信息与其所在的家庭、职业及邻里关系进行割裂的处理方法，也会使得我们无法触及我们真正希望去理解的行为及特征的本源。同理，如果将一个组织与其成员以及该组织之上更大的系统割裂开来，也必将限制我们充分地理解组织的能力及作用(Beatty, Harris & Barnes, 2010; Harris, Luke, Burke & Mueller, 2008; Luke et al., 2010)。

指数随机图模型(ERGM)是一种专门针对关系数据的统计方法，本书的目的就在于为指数随机图建模提供一种相对非技术性的介绍。正如在图 1.1 和图 1.2 中所展示的那些工具一样，指数随机图模型是一种用于识别并检验关系模式的工具，它能够识别出观测网络中存在哪些关系模式、网络成员或者社会力量所具有的哪些特征是具有解释力的。指数随机图模型是一种独特且有效的网络统计工具，因为它能够通过类似逻辑回归(logistic regression)的统计形式来解释其所观测到的网络结构特征。确切地说，指数随机图模型可以被用于理解一个观测网络的形成是源于网络成员某种属性特征(如年龄以及就业情况等特征)还是源于网络形成过

程中的关系模型。

到目前为止，大多数关于指数随机图模型的文章都是由统计学家撰写的，而对于应用社会科学家而言，这些文章不太容易理解。为了填补这一空白，本书的目标阅读群体是那些从事社会科学领域研究的教师、科研人员以及研究生。我们希望读者通过阅读本书，能够建立、评价并解释一个复杂的指数随机图模型。

本章余下的部分包含一个简要的历史介绍以及对于重要网络概念及词汇的概括。第 2 章则概要式地介绍了指数随机图模型的历史发展、理论、形式以及特征。具体的指数随机图建模过程将会在第 3 章和第 4 章中展示。其中，第 3 章展示了一个复杂的指数随机图模型从始至终的构建过程，该章包括了探索性网络分析、模型估计、解释、诊断以及拟合优度评价等方法。第 4 章将进一步讨论指数随机图模型的模型估计和解释、针对有向网络的模型拟合问题，以及利用二元组协变量和其他网络作为模型的自变量的问题。最终，第 5 章概括式地列举了全书中所采用的方法以及相关的外部学习资源。需要注意的是，本章中的叙述并不是对于社会网络分析（social network analysis，SNA）的一个介绍，虽然本书后续的部分也涵盖了社会网络分析的历史以及术语等相关内容，而且在第 3 章的开始部分（作为网络模型构建过程的一部分）包含了一些基础网络分析和可视化工具介绍，但这两点是需要区分开来的。

第1节 | 历史与概念

尽管在18世纪乃至更早期已经有一些对于社会关系和社会网络的描述(Buchanan, 2002; Caulkins, 1981; Freeman, 1996; Luke & Harris, 2007),但约瑟夫·莫雷诺于1934年出版的著作《谁将生存?》(1953后再版)被认为是现代社会网络分析的开端(Freeman, 1996; Wasserman & Faust, 1994)。莫雷诺在他的书中描述了一种新的对于关系的表述方法:利用一系列的点来表示人,利用点之间的连线来表示关系。这种表示方式被称为社群图,也是最早的社会网络分析工具。自莫雷诺的开创性工作之后,社会网络分析经过几代的发展逐步兴盛起来(Wasserman & Pattison, 1996)。与传统的统计方法要求观察样本之间彼此独立假定不同,现代的社会网络分析具有以下四项主要特征(Freeman, 2004, 2011):

(1) 它根源于社会行动者之间的关系及在此之上的结构性思想;

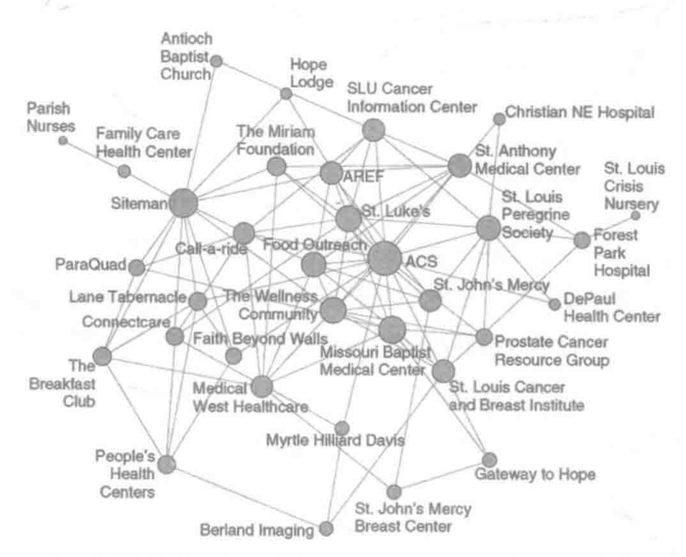
(2) 它是以系统的实证数据为基础的;

(3) 它非常重视关系图形的表象功能;

(4) 它依赖于数学或计算模型的使用。

网络可以由个体、组织、事件、出版论文以及相互连接的任何事物构成。学校中的学生以及他们之间的朋友关系可

以构成一个网络，与此相仿，一个城市的管理者以及他们共同所属的理事会之间的关系，或者是一群正在策划发动恐怖袭击的恐怖分子之间的关系，都可以构成一个网络。网络不仅可以由人构成，例如，图 1.3 中就展示了一个医疗服务网络，在该网络中，癌症治疗机构之间通过相互协作为城市中缺医少药的病人提供服务。虽然自莫雷诺时代以来网络理论已经取得了长足的进展，然而，不难发现，图 1.3 所示网络的可视化形式与莫雷诺所绘制社群图具有惊人的相似度。在图 1.3 中，每一个圆圈，或者说节点代表了一个组织，而连接两个圆圈之间的连线则表示了两个相互联系的组织之间的沟通关系。节点的规模显示了该组织在网络中具有联系的数量；而节点的标签则明确地标明每一个节点所代表的组织名称。



该网络展示了在一个城市环境中，为了救护缺医少药的癌症病人，各医疗组织之间是如何进行沟通的(Harris et al., 2011)。

图 1.3

上述网络中有两个特征是显著的，而且，这两个特征将会成为本书后续部分反复提及的内容。首先，网络成员所具有的联系并不是均匀分布的。即网络中的机构与其他机构建立联系的频次并不是均等的，有些机构仅会联系一家或者两三家机构，而有一些机构则会与网络中的其他机构建立广泛的联系。其次，居于网络中心位置的一群机构，利用群体间业已建立的广泛联系形成了一个紧密联系的聚簇。上述两项社会网络的基本结构特征——网络成员所具有的联系数量分布以及网络中存在着紧密联系群体，正是网络科学家不断努力力求能够更准确解释和表现的网络特质。

图 1.3 中简化的网络图形掩盖了潜在数据的复杂性，这种复杂性一直阻碍着科学家们采用统计建模的方法，而统计建模的方法能对观测网络进行更精确的预测和解释。因此，实证网络研究中几乎就回避了使用统计建模的方法（Hunter & Handcock, 2006; Snijders, 2011a），并主要采用网络图形可视化的视觉检验方法以及描述性统计方法来识别关系模式，并对关系特征进行描述（Shumate & Palazzolo, 2010）。这种回避的做法极大限制了研究人员从社会网络分析中获得有效分析结论。例如，在梅毒传播网络中（参见图 1.2），基于网络的描述性统计分析以及可视化方法都察觉到，受感染者之间存在的聚集现象以及高致病概率的性传播模式，与网络成员的性别、种族以及药品使用情形之间可能存在关联关系。然而，上述方法不具备对这些假设进行检验的能力：这些关系模式与随机产生的现象是否不同，而事实上随机属性或多或少对于解释高致病概率的性传播模式具有意义，或是整体网络的属性（如呈现非均匀状态的关系分布）如何解释

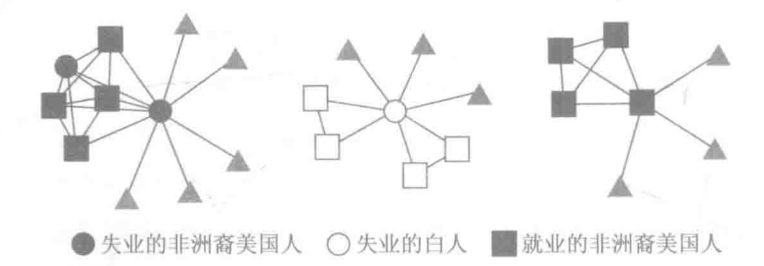
网络中心所出现的性关系聚集现象。

这种对于统计网络模型采取回避的做法在最近的10到20年时间里发生了巨大的变化。统计网络模型出现了两个分支(van Duijn & Huisman, 2011)。第一个分支关注网络中的行动者,而另一个分支则关注网络中的关系。以行动者为中心(actor-focused)的统计网络模型由一系列区分行动者群体以及试图解释或预测行动者属性的模型构成。而以关系为中心(tie-focused)的分支则是由一系列旨在解释或者预测关系形成以及关系模式的模型构成。下面这一部分将描述几个统计模型,这些统计模型在两个统计网络模型分支上均有所体现。当然,这不是一个详尽的列表,还有一些面向具体情形的网络模型没有涵盖进来(Scott & Carrington, 2011)。

在以行动者为中心的分支中,最简单的方法是将网络信息作为独立自变量纳入到标准的广义线性模型以及非参数模型中。例如,一个以行动者为中心的研究方法如果选择对长寿问题进行分析,将会选择美国加利福尼亚州阿拉米达县有代表性的成年人做为样本,采用卡方(chi-squared)检验的方法来判断社会网络特征(如,亲密朋友的数量)与死亡率之间的关系。这种特殊的研究方法检验的是自我中心网的数据,该数据不同于之前我们提到的癌症组织网络中所用到的数据(属性数据)。

自我中心网络(egocentric network)的数据通常是来源于个体对他们相关个人网络的描述。例如,图1.4的三个自我中心网就是一项针对失业问题的研究,在该项研究中,39位求职者被问及在最近找工作的过程中所联系的人和机构(Harris & Baker et al., 2012)。其中,每一个自我中心网络

都是一个从个体视角出发的寻觅工作机会的网络，而整个数据集一共包含了 39 个这样的网络。每一个网络正中心的图形就是自我中心(ego)，或者称之为参与者(participant)；其他的图形则代表了该自我中心在最近找工作的过程中所联系的组织和人员。在图 1.4 中，阴影部分用来区分种族，形状则用来表示就业状态。这种以行动者为中心的研究方法旨在理解，如何通过这些自我中心网络来解释处于不同就业状态人员所表现出来的在行动者属性特征上存在的差异？尤其是，求职者的自我中心网的规模和组成如何解释他(或她)所处的就业状态？



从三个参与者的求职自我中心网视角观察失业问题(Harris & Baker et al., 2012)。

图 1.4

不是所有以行动者为中心的模型都使用了自我中心网数据，有一些采用的是整体网络(whole network)数据。整体网络数据通常是通过定义一个研究者关心的网络(例如，城市中提供癌症治疗服务的组织)，并找出符合该定义的一系列网络成员名单，从而建立整体网络的。这里，网络成员通常会被问及他们与成员列表名单中其他成员的关系，从而形

成一个单一的、能够展现所有网络成员之间关系的网络。前面提到的癌症组织网络就是一个整体网络的例子。该网络是通过如下方式来测量的：首先，识别一个城市范围内为缺医少药的癌症患者提供服务的机构名录，然后，调查每一个机构与其他所有机构之间的关联。

除了使用标准的统计模型方法（例如卡方检验）之外，还有一些以行动者为中心的建模方法。例如，与潜类分析(latent class analysis)相似，随机块模型(stochastic blockmodels)方法将个体行动者根据近似相同的原则划分为若干块或者不同位置，于是，网络可以通过若干块之间的关系来表示(Anderson, Wasserman & Faust, 1992)。以行动者为中心方法还包括传染病模型(contagion models)，该模型旨在通过空间回归方法(spatial regression approach)将一个网络视为一个预测变量(van Duijn & Huisman, 2011)。另外，还有纵向网络模型(longitudinal network models)或者社会网络动力学(social network dynamics)的统计模型，这些模型既可以是以行动者为中心的模型也可以是以关系为中心的模型，这些模型将随时间而产生的网络变化视为行动者的选择引入模型中来，或者将网络变化视为网络中二元组或关系层次的变化函数(Snijders, 2002, 2011b; Snijders, van de Bunt & Steglich, 2010)。最近，以行动者为中心的方法还引入了潜在位置聚类模型(latent position cluster model)。这种复杂的模型可以将行动者的属性（例如，年龄、性别、种族）整合进潜在空间的网络成员间的聚类算法中(Krivitsky & Handcock, 2008; van Duijn & Huisman, 2011)，从而识别差异化的聚类模式。更多的有关以行动者为中心的模型信息，包括专门对

不同类型行动者模型的资料，可以参考范・杜因和哈斯曼(van Duijn & Huisman, 2011)的文章。

以关系为中心的研究分支旨在解释或分析关系以及关系的模式，该分支包括了一系列针对多种数据类型及不同研究问题的模型。其中，二次指派程序(quadratic assignment procedure, QAP)和多元回归二次指派程序(multiple regression quadratic assignment procedure, MR-QAP)都可以用来测量网络之间的相关性(Krackhardt, 1987; van Duijn & Huisman, 2011)。社会关系模型(social relations model, SRM)是一种由嵌入在行动者中的二元组构成的多层次回归模型(multilevel regression model)，这里模型中的行动者是作为一个联系的发送者或者接受者而言的(Kenny & La Voie, 1984; van Duijn & Huisman, 2011)。社会关系模型可以用来构建包含加权关系的复杂网络模型，还可以通过增加网络层次的方式将模型扩展到多层次网络模型。然而，上述方法往往仅针对一种特征模型或问题。指数随机图模型(ERGM)虽然也是以关系为中心的方法，但它与之前所讨论的其他以关系为中心的方法不同，它是以概率分布的指数族为基础的统计模型大家族中的一部分。在这个指数族中包括许多模型，其中，大多数模型都是针对非网络数据的。然而，指数随机图模型则可以专门针对网络数据，也可以包含网络成员的属性数据，甚至可以将整体网络特征作为自变量来预测二元网络的网络结构问题(二元网络是指网络中的连线仅用0和1来表示是否存在联系，连线本身并没有具体权值)。最终，学者们正在努力将现有的估测复杂指数随机图模型的程序扩展到包括加权网络(Krivitsky, 2012)。

第 2 节 | 网络术语

> 任何新人都可能会在(网络分析的)术语丛林中种植下一棵树,这恰恰彰显了网络分析方法易于被人理解与接受的特质。
>
> ——巴恩斯(Barnes, 1972:3)

自 20 世纪 70 年代以来,网络科学经历了相当快速的发展。正如巴恩斯在 1972 年所指出的,在网络科学领域一个术语丛林已经形成,这里有许多术语其实是一个意思,而另外有一些术语则可能同时包含很多不同的意思。如图 1.5 所示,我们将对本书中所用到的一些网络概念和定义进行简单的描述,还有一些术语我们将根据需要在书中内容部分以及附录中进行解释。

一个网络是由一系列行动者(actors)及其关系(relationships)所构成的。行动者也被称为节点(nodes)、顶点(vertices)、个体(individuals)或者成员(members)。而行动者之间的关系又经常被称为链接(links)、线条(lines)、联系(relations)或者关系(ties)。这些链接、线条、关系或者联系可以是有方向的也可以是无方向的;可以是二分数据(存在/缺失)也可以是有值/加权数据;如果两个行动者之间的关系

存在方向，虽然本书中自始至终没有采用该种数据类型（例如，卡伦送钱给彼得），那么，行动者之间的关系就是有向的，也经常被称为弧（arc），利用一个箭头表述，例如卡伦→彼得。有向关系可能是单一方向形式（非对称），也可能是双向形式（相互或互惠），这种双向形式可以利用一个双向箭头来表示，例如，卡伦↔彼得。如果两个行动者之间的关系并没有一个专指的方向（例如，卡伦和彼得共进午餐），这种情况（连线）通常被称为边（edge），用一条连接两个行动者之间的连线（卡伦—彼得）来表示（没有箭头）。

一个网络中所有成对的节点都是二元组（dyads）。二元组可能相互连接，也可能不连接。在一个有向网络中，二元组可以通过一个非对称的联系或者交互的联系建立起连接。三元组（triads）则是网络中三个节点的子集（sets）。同二元组一样，这些三元组之间既可以是连接的，也可以是不连接的。图 1.5 包括二元组、三元组以及其他一些网络重要的特征的图形展示形式。

本书中的剩余部分将关注解释和阐述指数随机图模型。指数随机图模型包含了以联系为中心的统计网络模型，就如同图 1.3 中的癌症组织网络一样。这里，联系是指一个二分变量（存在或不存在），而网络信息来源于对整体网络的横截面信息的提取。如果在进行网络分析的过程中发现上述描述方法并不合适，可以参考一些其他资料，例如，沃瑟曼和福斯特（Wasserman & Faust，1994）、斯科特和卡林顿（Scott & Carrington，2011），以及瓦伦特（Valente，2010）。书中的第 2 章描述了指数随机图模型的历史发展以及统计理论基础；第 3 章和第 4 章展示了一个复杂的指数随机图模型从开始

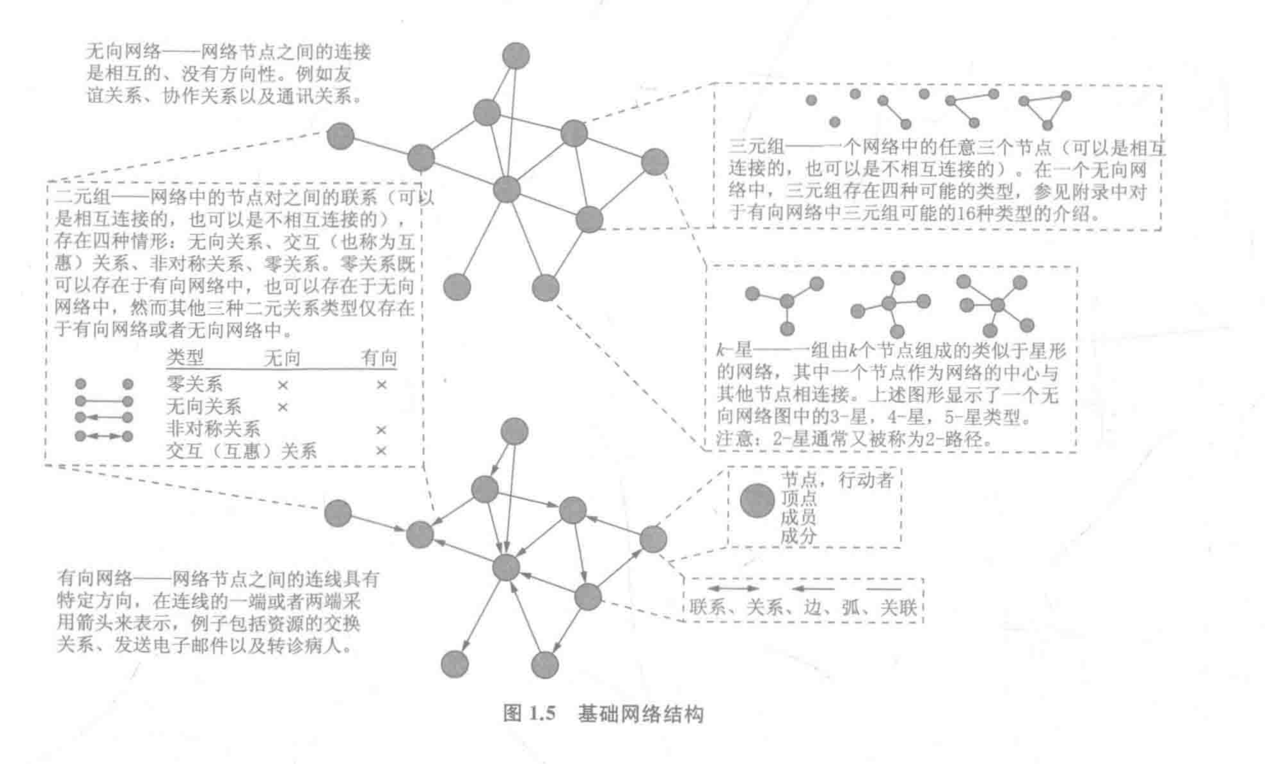

无向网络——网络节点之间的连接是相互的、没有方向性。例如友谊关系、协作关系以及通讯关系。
三元组——一个网络中的任意三个节点（可以是相互连接的，也可以是不相互连接的）。在一个无向网络中，三元组存在四种可能的类型，参见附录中对于有向网络中三元组可能的16种类型的介绍。
二元组——网络中的节点对之间的联系（可以是相互连接的，也可以是不相互连接的），存在四种情形：无向关系、交互（也称为互惠）关系、非对称关系、零关系。零关系既可以存在于有向网络中，也可以存在于无向网络中，然而其他三种二元关系类型仅存在于有向网络或者无向网络中。
类型　无向　有向
零关系　×　×
无向关系　×
非对称关系　×
交互（互惠）关系　×
k-星——一组由k个节点组成的类似于星形的网络，其中一个节点作为网络的中心与其他节点相连接。上述图形显示了一个无向网络图中的3-星，4-星，5-星类型。
注意：2-星通常又被称为2-路径。
节点，行动者
顶点
成员
成分
联系、关系、边、弧、关联
有向网络——网络节点之间的连线具有特定方向，在连线的一端或者两端采用箭头来表示，例子包括资源的交换关系、发送电子邮件以及转诊病人。

图 1.5　基础网络结构

到结束的构建过程，本书并没有包含更广义的指数随机图模型——考虑加权关系的模型，目前，对于广义指数随机图模型的研究仍在进行中（参见 Krivitsky，2012）。

第2章

统计网络模型

对于整体网络的观察有助于我们认识社会力量(social forces)是如何塑造真实世界系统的。建立以关系为中心、面向整体网络分析模型的目的在于:解释并预测网络中两个行动者之间关系的形成。复杂的以关系为中心的整体网络模型可能包含网络成员的属性特征、网络全局结构特征等因素,最终,通过结合这些因素构建的复杂模型可以用来解释并预测网络关系的形成。这些模型可以分为以下四种类型:

(1) 简单随机图模型(simple random graph models);

(2) 二元独立性模型(dyadic independence models);

(3) 二元依赖性模型(dyadic dependence models);

(4) 高序依赖性模型(higher-order dependence models)(Robins, 2011; Wasserman & Robins, 2005)

上面所列举模型是依据其不断增加的复杂程度进行排序的,这个顺序也恰恰反映了以关系为中心的统计网络模型的发展历程。随着不断添加更为复杂的假设条件,统计网络模型不断完善。本章有两个平行的目标:(1)描述这四类模型的发展历程;(2)审视这四类模型的统计形式及其所依据的假设。统计建模,尤其是针对网络的统计建模,其目的在

于解释观测网络与随机发生网络之间的差异，因此，在讨论更复杂模型的构建及形式之前，我们首先介绍简单随机图(simple random graphs)模型。

第 1 节 | 简单随机图

在审视整个模型发展过程之前，将简单随机图的特征作为理解整个网络模型的基础十分有必要。一个简单随机图是在由 n 个节点构成的所有可能网络中随机选择的网络，其中，网络中的每一条连线(联系两个节点)都以同样的特征概率发生(Frank & Strauss，1986；Karonski，1982)。因此，一个随机图中节点之间的关系都是基于某种概率随机发生的，关系之间是相互独立的。即，在一个包含同事之间朋友关系的网络中，道格和金姆成为朋友的概率是独立于网络中其他朋友关系的，包括道格和金姆与网络中其他成员之间成为朋友关系的概率；同时，道格和金姆成为朋友关系的概率与网络中其他朋友关系形成的概率是一致的。因此，一个简单随机图模型就是在网络成员之间随机分配关系，不考虑网络成员的属性影响因素以及任何可能影响关系形成的社会力量。

简单随机图中关系发生的概率是网络中所观测的关系数占所有可能的关系数的比例。计算所观测的关系数占所有可能的关系数的比例与网络密度(network density)的概念是一致的，该概念显示了网络联系的密集程度，是通过从 0(表示网络成员之间完全不存在联系)到 1(表示网络中所有可能的联系都实际存在)之间的范围来评价的。对于一个无向网络而

言，网络密度可以采用如下公式来计算：$\frac{L}{n(n-1)/2}$，其中，L 是网络中边的数量，n 是网络中节点的数量，而网络中节点的数量(n)就是网络规模(network size)。

为了描述简单随机图和所观测的真实网络之间在结构特征上的差异，图 2.1 展示了一个简单随机图，该图与图 1.3 中所显示的癌症组织网络具有同样的网络规模和密度。通过对简单随机图与观测网络进行比较可以帮助我们认识：观测网络的某些特征并没有出现在简单随机图中。这种类似比较对于统计网络模型构建是十分关键的。

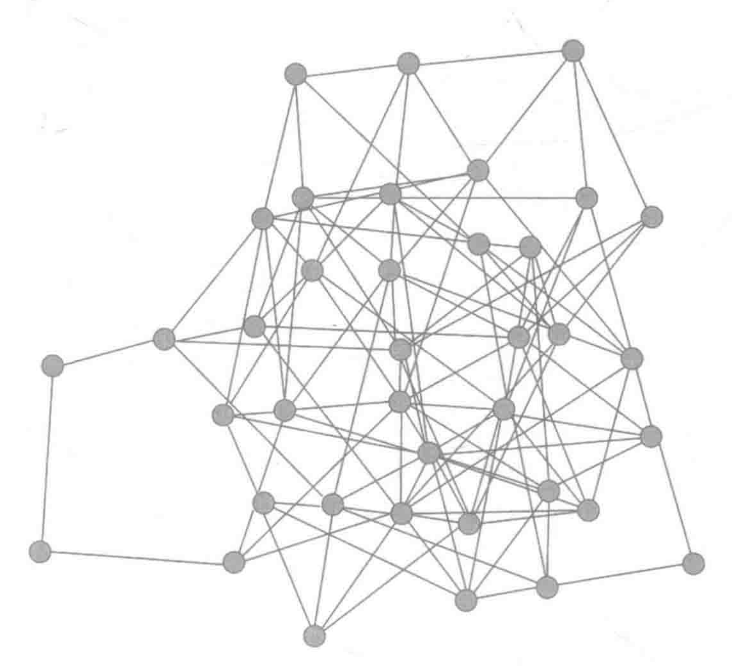


图 2.1 一个包含 38 个节点和任意两点间关系形成概率为 14%的简单随机图

虽然各种社会力量均对网络的形成产生了影响，而且影响之间还存在较大的差异，但是网络科学家已经发现，在真实的观测网络中存在许多结构性特征（Snijders, 2011a; Rivera, Soderstrom & Uzzi, 2010），这些特征能够将观测网络与简单随机图显著区别开来：

1. 网络成员在建立关系的倾向上并不是完全相同的，即非均匀性度分布（nonuniform degree distribution）。例如，在一个工作场合的朋友关系网络中，有一部分工人与其他工人相比，他们能够获得更多的朋友关系。

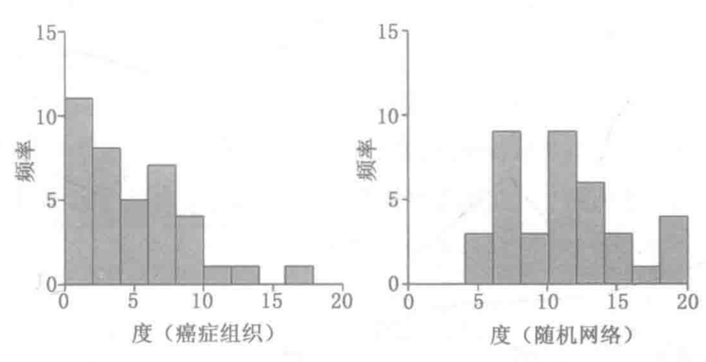
2. 具有相似特征的行动者之间建立联系的概率往往要高于基于随机联系产生的概率，即同质性（homophily）。在一个学校的环境下，如果布莱恩和本都是男孩，那么他们之间成为朋友关系的概率就会大于随机选择的学生之间成为朋友关系的概率，而贝琪可能与南茜交朋友的原因恰恰是她们都是女孩。

3. "朋友的朋友也是我的朋友"发生的概率通常要高于随机发生的概率，即传递性（transitivity）。如果，南茜是布莱恩的朋友，而布莱恩又是本的朋友，那么，南茜和本就很有可能成为朋友。

4. 真实有向网络往往会比随机（有向）网络产生更多的互惠性（reciprocity）联系。当布莱恩传送消息给贝琪时，贝琪也会给布莱恩传送信息，而且这种回馈的概率要比随机期望的概率要更高（布莱恩↔贝琪）。

通过观察比较图 2.1 简单随机图与图 1.3 的癌症组织所形成关系网络之间的差异，能够很好地诠释上述提到的诸多差异中的两项差异：非正态的度分布以及传递性。癌症组织

网络的度分布呈现明显的右偏倾向，少量的组织拥有大量的联系，而在简单随机图中的度分布则接近于均匀分布(参见图2.2)。实际上，一个不断衰退的度分布恰恰显示了大多数的网络成员仅拥有少量的联系，而少数成员则拥有大量的联系。这种不断衰退的分布特征不仅出现在癌症组织网络中，它也体现在许多具体的观测网络中。



通过直方图展示：具有相同的节点规模($n=38$)以及边密度($d=0.14$)特征的癌症组织网络(左)和简单随机图网络(右)在度分布上的差异。

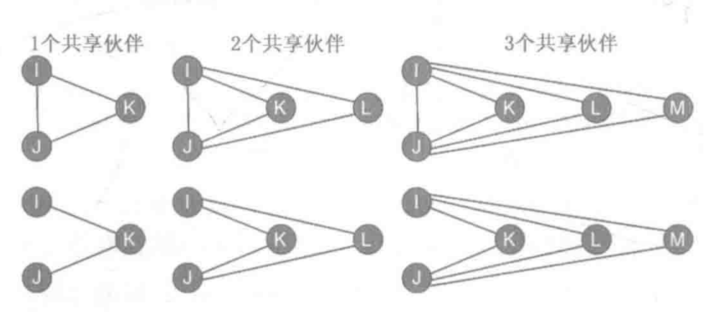
图 2.2

在无向网络中，传递性或者说是“朋友的朋友也是我的朋友”的属性，可以通过观察完全三元组子结构(即由三条边构成三角形)或者三元组中三种边自由组合的情况(参见图1.5对三元组的图形描述)来实现。癌症组织网络中共有71个三角形，而在简单随机图网络中则仅有32个三角形。

除了三角形之外，网络中其他两种特征结构的存在情况也可以帮助我们区分三元传递结构是否存在：边共享伙伴(ESP)以及二元组共享伙伴(DSP)。边共享伙伴是指存在一个连通的二元组(参见图1.5中的二元组类型)，且该二元组

的每一个成员都与网络中的第三方成员相连接的情形(参见图 2.3)。在任意给定的三角组结构中,都存在三种边共享伙伴情形。例如,图 2.3 的左上角所显示的三角形中,IJ 的边共享伙伴为 K,IK 的边共享伙伴是 J,KJ 的边共享伙伴是 I。网络中的边共享伙伴的数量是三角形数量的三倍。每条边也可以共享超过一个伙伴(图 2.3),边存在多个边共享伙伴是网络中紧密连通聚类特征的一种表现。

一个二元组共享伙伴(DSP)是指一个二元组(无论连通与否),二元组的每一个成员都与网络中的第三方成员相连接。二元组共享伙伴被认为是传递性的先决条件,因为一个没有连通的二元组共享伙伴仅需要增加一条边就可以完成一个或者多个三角形的构造过程。

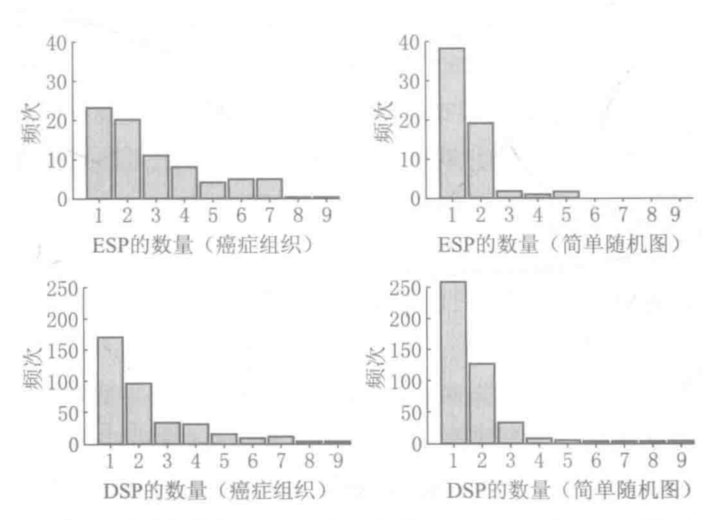


二元组 IJ 的边共享伙伴(图上方)以及二元组共享伙伴(图下方)。其实,边共享伙伴也可以被认为是一种特殊的二元组共享伙伴类型。

图 2.3

网络中边共享伙伴的分布显示了有多少相互连通的二元组,以及这些连通的二元组具有 1 个共享伙伴、2 个共享伙伴等的情形(如图 2.4)。同样的,二元组共享伙伴的分布也显示了网络中具有 1 个共享伙伴、2 个共享伙伴等情形的二

元组的数量。观测网络往往较随机网络在二元组共享伙伴的数量上具有优势(参见图 2.4),这显示了观测网络要比随机网络具有更强的传递性以及预传递结构。图 2.4 中的条形图显示了:在实际观测的癌症网络中,有一些边拥有 5 个以上共享伙伴,而在随机网络中,大多数的边仅存在一个或者两个共享伙伴(图上方)。二元组共享伙伴也存在同样的情形(图下方)。



具有同等网络规模($n=38$)以及边密度($d=0.14$)的癌症组织网络(图左侧)和简单随机网络(图右侧),以及上述两个网络分别所对应的 ESP 分布(图上方)和 DSP 分布(图下方)。

图 2.4

第 2 节 | ERGM 的发展

以关系为中心的指数随机图模型(ERGM)家族的发展最早可以追溯到 1959 年简单随机图模型的提出。下面这一部分所描述的四种模型类型分别代表 ERGM 发展过程中四个关键的里程碑,分别为:简单随机图模型、二元独立性模型、二元依赖性模型,以及高序依赖性模型。

简单随机图模型

1959 年,艾多斯和瑞尼提出了一种利用统计模型来获取简单随机图特征的方法(Erdös & Renyi, 1959; Frank & Strauss, 1986; Karonski, 1982)。该模型的统计形式如下:

$$P(Y=y)=\left(\frac{1}{c}\right)\exp\{\theta L(y)\} \qquad [2.1]$$

这里 c 是一个常数,它确保了随机网络 y 的概率是在 0 到 1 之间,$L(y)$是网络 y 中边的数量,θ 是边条件项的系数。简单随机图模型所包含的一个假设是:网络成员之间关系的产生是随机的,是独立于其他成员之间关系的。简单随机图模型通常并不能够很好地把握所观测网络的结构特征,因为简单随机图模型方法忽视了对于影响网络关系形成的社会

力量的关注。虽然该模型无法在结构方面提供更多有用的信息，但它却提供了一条与其他更复杂模型进行比较以及未来进行模型改进评价的基线。

二元独立性模型

ERGM发展过程中的另一个重要阶段出现在1981年，霍兰德(Holland)与莱因哈特(Leinhardt)发展了一种针对二值有向网络的模型，该模型目标在于理解观测网络的两项特征：(1)观测网络的入度(indegree)分布存在较大差异，或者说行动者接受到链入关系的数量与其期望值之间存在较大的差异；(2)互惠(或者交互)关系的发生经常与期望值存在较大差异。霍兰德和莱因哈特提出利用 p_1 模型来估计由于互惠性(reciprocation)以及差异化吸引性(differential attractiveness)所造成的数量差异，并利用这些特征来检验一个观测网络存在的概率问题。p_1 模型是第一个可以针对网络规模与密度均存在差异的有向网络直接进行比较的模型。p_1 模型的统计形式包括四个部分，分别对应一个二元组(ij)四种可能的状态：(1)两点之间不连通，即 P_{00}；(2)从 i 到 j 之间存在非对称的连线，即 P_{10}；(3)从 j 到 i 之间存在非对称的连线，即 P_{01}；(4)在 i 到 j 之间存在一条互惠的连线，即 P_{11}。每一种状态的概率如下：

$$
\begin{aligned}
P_{00} &= e^{k_{ij}} \\
P_{10} &= e^{k_{ij}+\alpha_i+\beta_j+\mu} \\
P_{01} &= e^{k_{ij}+\alpha_j+\beta_i+\mu} \\
P_{11} &= e^{k_{ij}+\alpha_i+\beta_j+\alpha_j+\beta_i+2\mu+\rho_{ij}}
\end{aligned}
\qquad [2.2]
$$

这里，k 代表一个标准化常数，α 表示连接的发送方，β 则表示连接的接收方，μ 表示网络的密度，ρ 代表网络中互惠连线的数量（An，2011；Holland & Leinhardt，1981；Van Duijn，Snijders & Zijlstra，2004）。基于上述四种条件，所观测网络 y 的概率可以表示为（An，2011）：

$$P(Y=y) \propto \exp\left[\mu L(y) + \sum_{i}^{N} \alpha_i y_{i+} + \sum_{j}^{N} \beta_j y_{+j} + \rho M(y)\right] \tag{2.3}$$

这里，$L(y)$是网络中连接的数量，y_{i+} 是网络中连出关系的数量，y_{+j} 是网络中链入关系的数量，而 $M(y)$是网络中具有互惠关系的数量。然而，p_1 模型较简单随机图模型更有意义的地方在于，该模型所依赖的假定是二元独立性假设，或者说是假定二元组在统计上具有独立性。正是基于这种假定，p_1 模型无法解释传递性、派系以及除互惠性和差异化吸引性以外的一些网络结构特征问题（Holland & Leinhardt，1981）。霍兰德与莱因哈特（1981）认识到了该模型的局限性，并利用这种局限性为该模型命名。p_1 模型中的下标 1 表示 p_1 模型是一系列模型中的第一个。他们工作的目标就在于通过建立一系列模型来推动理论与方法的进步，从而使模型能够包含更加复杂的依赖性假设。

虽然在后续的原创性实证分析中，p_1 模型并没有被广泛采用，但该模型确实成为后续模型构建过程中若干主要基石之一。霍兰德与莱因哈特的若干贡献中的其中一项就是推进了利用分布的指数家族来判断观测网络的概率，这一点在 ERGM 名字中就有所体现。指数家族就是一组概率分

布，包括：正态分布、卡方分布、指数分布以及其他通常所使用的分布。更进一步来说，任何像公式 2.4 这样的具有一个概率密度的随机变量都来源于指数家族。

$$f(x) \propto \exp\{\theta' s(x)\} \qquad [2.4]$$

二元依赖性模型

ERGM 发展过程中另一个具有里程碑意义的事件发生在 5 年之后的 1986 年。弗兰克（Frank）和斯特劳斯（Strauss）利用指数家族分布，整合了网络成员之间的依赖性假设，确立了新一代的统计网络模型。进一步而言，弗兰克和斯特劳斯的模型将马尔科夫依赖性假设（Markov dependence assumption）引入到模型中来，该假设强调了那些没有共享节点的边之间的条件独立性（Frank & Strauss，1986）。即假定网络中的其他因素不变时，行动者 A 与行动者 B 之间建立联系的概率与行动者 C 和行动者 D 之间建立联系的概率是彼此独立的，因为这两项联系不包含共同的行动者*。马尔科夫依赖是一种广义的假设，即假定包含同一个节点的若干连线之间是相互依赖的，这是一种二元依赖性模型的形式。举例而言，在一个交友网络中，克林特和利安娜的友谊关系与利安娜和埃文的友谊关系之间存在相互依赖关系，因为这两组朋友中都包含一个共同的成员——克林特。正是考虑到这种依赖性，弗兰克和斯特劳斯仅仅用一些统计项就提出了马尔科夫随机图模型：

---

* 这是对之前二元独立性模型假设的进一步约束。——译者注

$$P(Y=y)=\left(\frac{1}{c}\right)\exp\{\theta L(y)+\sigma_k S_k(y)+\cdots+\tau T(y)\}$$

[2.5]

这里，c、$L(y)$以及θ分别代表了常数、网络中连接的数量，以及连接数量的系数。这些统计项与简单随机图模型所使用的统计项是一致的(公式 2.1)。$S_k(y)$项则代表了网络中具有 k-星的子图数量(参见图 1.5 中 k-星的若干形态)；因此，$S_2(y)$也就代表具有 2-星特征子图的统计项。将 k-星项纳入模型中来的主要目的是为了帮助了解度分布的不均衡性，在网络中，无论是具有低度数的节点(例如，1-星节点、2-星节点、3-星节点)还是具有高度数的节点(例如 4-星节点、5-星节点、10-星节点)，它们都被视为一个独立的 k-星项。对应的 σ 是对于每一个 k-星项的参数估计值。$T(y)$项则表示网络中三角形的数量，τ 则是其对应的参数。与简单随机图模型相比，弗兰克和斯特劳斯的模型将观测网络中更多的网络结构特征纳入到了模型中，但没有将网络成员特征(network member characteristics)纳入统计网络模型框架中作为协变量，例如性别和种族等。协变量是社会过程(social process)中十分重要的因素，因此，协变量的缺失限制了该模型的应用。

弗兰克和斯特劳斯模型作为后续的统计网络模型发展(Wasserman & Pattision, 1996)的基础具有极为重要的作用——沃瑟曼和帕蒂森在 10 年后扩展了马尔科夫模型。为了与霍兰德和莱因哈特(1981)的 p_1 模型相区别，沃瑟曼和帕蒂森将他们新的依赖性模型命名为 p^* (p-星)模型(Hunter & Handcock, 2006)。p^* 模型假定在网络中各连线

之间存在更广泛的条件依赖关系，具体而言就是，当网络中其他连线已经确定条件下，两条具体的连线之间存在条件依赖关系，且两条连线同时存在的条件概率不等于两条连线各自边际条件概率（marginal conditional probabilities）的乘积（Wasserman & Pattison，1996）。也就是说，任意两条连线同时存在的概率是不同于这些连线各自存在概率组合的。这一假设的普遍性使得我们可以将含有大部分依赖性假设条件的 ERGM 看作是 p^* 模型。p^* 模型也具有了整合协变量的能力。p^* 模型的一般形式表示为：

$$P(Y=y)=\left(\frac{1}{c}\right)\exp\left\{\sum_{k=1}^{k}\theta_k z_k(y)\right\} \qquad [2.6]$$

这里，$\frac{1}{c}$ 是一个常量，用来确保概率始终保持在 0 到 1 的范围之内，同时保证所有可能网络的概率和为 1。θ_k 是网络统计量[由 $z_k(y)$ 所表示]所对应统计项的参数。为了便于使用与解释，公式 2.6 中所表示的 p^* 模型的一般形式实际上就包含了公式 2.7 所示的模型。公式 2.7 并不是用来预测整个网络出现的可能性的，而是可以在网络中其他连线已经确定的条件下，预测一条连线出现的概率（Hunter，Handcock，Butts，Goodreau & Morris，2008）。还有一些关于公式 2.6 和公式 2.7 之间相关的技术信息可以参考亨特和汉考特等的文章（Hunter & Handcock et al.，2008）。

$$\text{logit}(P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c))=\sum_{k=1}^{\kappa}\theta_k \delta_{z_k(y)} \qquad [2.7]$$

在公式 2.7 的右侧，k 代表网络统计量的个数，κ 则代表

总数量。θ_k 表示每一项网络统计项所对应的系数。$\delta_{z_i(y)}$ 表示当增加一条 i 和 j 之间的连线时，即随着 Y_{ij} 从 0 到 1 变化时，网络统计量所发生的变化(Goodreau, Kitts & Morris, 2009)。这个 δ 统计量被称为变化统计(change statistic)，是逻辑回归模型(logistic regression model)与 p^* 模型在解释方面最关键的差异。在公式 2.7 的左侧，分隔符号"|"显示 $Y_{ij}=1$ 的概率是以剩余网络为条件的，这里 Y_{ij}^c 表示网络中除去 Y_{ij} 之外的所有二元组关系。和逻辑模型一样，分对数转换(logit transformation)被用于重新构建公式 2.7，使得公式的左侧成为以一条连线为条件更简化条件概率形式，这样更便于模型解释(公式 2.8)。

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=\text{logistic}(\theta_1\delta_{z_1(y)}+\theta_2\delta_{z_2(y)}\cdots) \quad [2.8]$$

公式 2.8 中所显示的 p^* 模型的表示形式和传统逻辑回归模型的使用和解释十分相像，除了两个重要的区别。第一个区别是：与系数 θ 相乘是公式中对应的变化统计量 δ，相比较而言，逻辑回归模型中与系数 θ 相乘的项通常是对应自变量的值。对于 p^* 模型中所整合的许多统计项而言，变化统计的结果是比较容易界定、识别并使用的；但对于有些涉及复杂依赖性(如传递性)的统计项，计算并解释这种变化统计结果时就会存在困难。第二，公式 2.8 右侧的分隔符号"|"之后的部分强调：当且仅当网络中其他因素都保持不变时，公式 2.8 右侧所表示的概率才适用。所以，以该模型为基础所计算的概率必须理解为：在网络中其他因素保持不变时，网络中 i 和 j 之间的连线的概率。

需要注意的是，p^* 模型中所考虑统计项不仅限于针对网络特征的统计项，就像简单随机图、p_1 模型以及弗兰克和斯特劳斯模型中的统计项那样，还包括针对网络成员的属性特征（例如，性别、年龄、收入等）的统计项。因此，p^* 模型就具有考虑网络成员属性特征的能力，这种属性特征经常是伴随着观测网络的结构特征存在的：互惠性、同质性、传递性以及非均衡的度分布。p^* 模型所具有的灵活性以及与逻辑回归模型相似的统计形式，使得 p^* 模型成为了社会科学家在进行网络研究时极为有效的工具。具体而言，p^* 模型可以用来理解一个给定实证网络是否可以由网络成员本地属性特征（例如，年龄、受雇状态等）以及整体网络的结构特征（如，传递性的数量）而形成。最后，虽然公式 2.6 和公式 2.7 是作为 p^* 模型提出的，但这些公式可以表示那些不符合 p^* 模型条件依赖假设的 ERGM 模型。例如，这种模型形式可以用来表示一个简单随机图模型，而后者的基本假设是：网络成员间的关系是独立的。

高序依赖性模型

尽管 p^* 模型不断地将传递性、同质性以及其他观测网络的特征统计项整合进模型中来，模型的灵活性得到了提升，但 p^* 模型仍然存在诸如近似退化（degeneracy）等问题，表明 p^* 模型仍无法充分获取观测网络结构特征。网络模型中的近似退化现象通常由这样的模型来表示，即其产生的模拟网络要么大部分为空图要么大部分全图（参见 Robins, Snijders, Wang, Handcock & Pattison, 2007，该文中图 1 是

一个极为经典的图形示例)。当对图形的统计结果进行平均计算时,那些由绝大部分为空图或大部分全图所构成的网络似乎也获得了貌似合理的统计结果,这种现象的出现凸显了对于模型拟合效果(model fit)进行仔细检验的必要性。

为了阐述近似退化的问题,帕蒂森和罗宾斯(Pattison & Robins, 2002)提出了一个局部条件依赖(partial conditional dependence)的假设。局部条件依赖性假定:不共享节点的两条连线之间的依赖关系可以依据网络中其他关系的存在而产生。2006 年,斯尼德斯(Snijders)和他的同事(Robins, 2011; Snijders, Pattison, Robins & Handcock, 2006)提出了适合于 p^* 模型的新模型参量(new model specifications),这些参量的提出为解决由帕蒂森和罗宾斯(2002)提出的局部条件依赖问题提供了一种专门方案。具体而言,社交圈依赖(social circuit dependence)就是局部条件依赖的一种形式。当两条连线能够构成一个 4 元循环(4-cycle)网络时,这两条连线可用于展示这种社交圈依赖。图 2.5 中,根据社交圈依赖,连线 AC 和连线 BD 条件依赖于连线 AB 和连线 CD 的存在。

斯尼德斯和他的同事(2006)当时并没有提出要改变 p^* 模型的形式以解释这种新的依赖性假设,相反,他们提出应该在 p^* 模型估计的过程中增加三个非线性的统计项。设计这三个统计项是为了解释条件依赖:几何加权度分布(geometrically weighted degree distribution)或者是交替 k-星(alternating k-star),交替 k-三角形(alternating k-triangle),以及交替 k-2 路径(alternating k-twopath)(Snijders et al., 2006)。这些统计项后来又经亨特和汉考特等人的修改,具

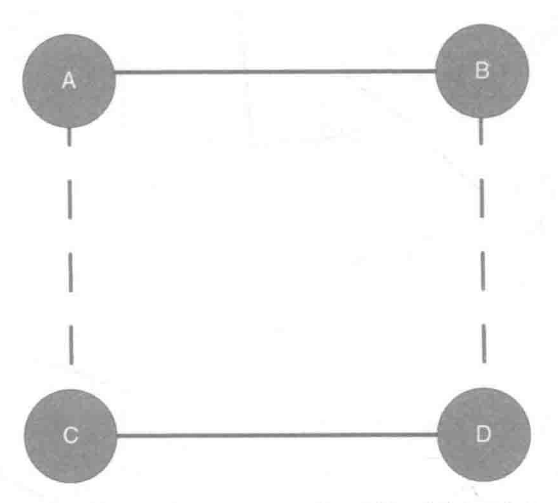


图 2.5 连线 AC 与连线 BD 之间的 4 元循环的社交圈依赖

有了更为简化的解释(Hunter，2007；Hunter & Handcock，2006)。具体而言，亨特和汉考特提出了几何加权度分布(geometrically weighted degree distribution，GWD)、几何加权边共享伙伴(geometrically weighted edgewise shared partners，GWESP)，以及几何加权二元组共享伙伴(geometrically weighted dyadwise shared partners，GWDSP)等统计项，作为斯尼德斯所解释观测网络中复杂结构以及依赖性条件的替换方法。这些经亨特和汉考特修改后的统计项将在下一部分中介绍；而原始统计项(新模型参量)则可以参见斯尼德斯和他的同事(2006)的论述。下面的论述主要遵循亨特和汉考特(2006)的相关定义。

几何加权度(GWD)。GWD 统计项被用于考察观测网络中不断递减的度分布特征(例如图 2.2 左侧)。该统计量是以每一个中心度值所对应的频数乘以一个加权参数然后求

和得到的：

$$u(y;\ \alpha)=e^{\alpha}\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^{i}\}D_{i}(y) \qquad [2.9]$$

这里 y 是代表一个网络，α 是一个所选择的或者估计的衰减参数(decay parameter)，i 代表中心度，而 $D_i(y)$ 则代表网络 y 中中心度为 i 的节点数量(Hunter, 2007)。$\{1-(1-e^{-\alpha})^{i}\}$这个乘数包含了几何函数，这个几何函数用于加权该网络统计量所对应的中心度；α 是一个中心度加权参数(degree weighting parameter)，用于控制权重。网络统计量的值 $u(y;\ \alpha)$是依赖于网络度分布以及选择的衰减参数 α 的。为了阐述 GWD 以及与之相对的一个度加权参数 α 值的计算过程，我们将使用图 1.3 和图 2.2 中的癌症组织网络作为示例。如图 2.2 所示，该网络具有一个不断衰减的度分布。表 2.1 显示了网络中不同度的值(i)，以及每一个 i 值所对应的频数，$D_i(y)$以及接下来对应三个不同 α 值的 GWD 计算结果。

表 2.1　对图 1.3 和图 2.2 中所显示的癌症组织网络进行 GWD 计算

Degree i	*Frequency* $D_i(y)$	$\{1-(1-e^{-\alpha})^{i}\}D_i(y)$		
		$\alpha=0.25$	$\alpha=0.5$	$\alpha=1.0$
0	4	0	0	0
1	2	1.56	1.21	0.74
2	5	4.76	4.23	3.00
3	4	3.96	3.76	2.99
4	4	3.99	3.90	3.36
5	5	5.00	4.95	4.50
6	0	0	0	0
7	3	3.00	3.00	2.88
8	4	4.00	4.00	3.90

续表

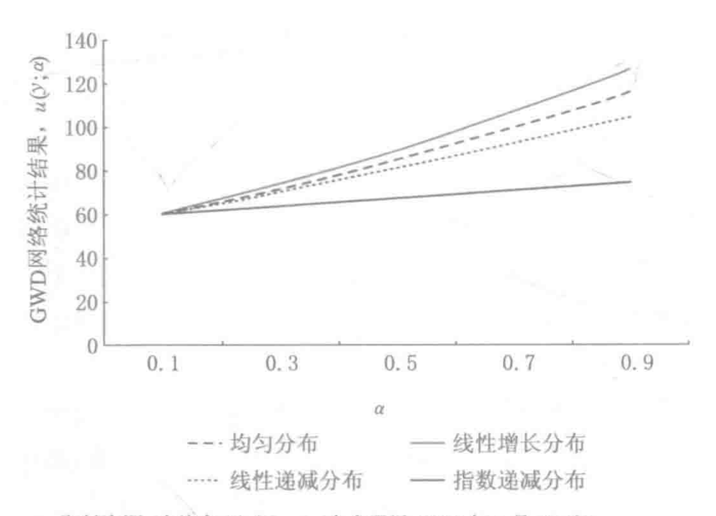
Degree i	Frequency $D_i(y)$	$\{1-(1-e^{-\alpha})\}D_i(y)$		
		$\alpha=0.25$	$\alpha=0.5$	$\alpha=1.0$
9	2	2.00	2.00	1.97
10	2	2.00	2.00	1.98
11	1	1.00	1.00	0.99
12	0	0	0	0
13	1	1.00	1.00	1.00
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	1	1.00	1.00	1.00
19	0	0	0	0
20	0	0	0	0
$\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^i\}D_i(y)$		33.26	32.04	28.30
$e^{\alpha}\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^i\}D_i(y)$		42.70	52.83	76.93

需要注意的是，在这些计算过程中有两个因素会影响到GWD统计的结果：网络中高度值节点的比例以及 α 值的选择。首先，通过将括号中的值增加到与相关的度的值一样，i 指数能够确保那些具有更高中心度的节点对整个网络统计结果产生更显著的影响。例如，一个度值为 13 的单一节点，其在括号中的值将被提升至 13 次幂。其次，$1-e^{-\alpha}$ 的值随着 α 增加而变大，而这些更大的值又会被提升至 i 的幂，高度节点由此再一次对整个网络统计产生更重要的影响。

观测网络通常服从不断衰减的度分布，因此，为了更好地理解 GWD 统计如何对观测网络的建模产生影响，检验不同 α 值下 GWD 值所对应的不同的度分布将是十分有益的尝试。图 2.6 展示了一个拥有 55 个节点的网络分别对应四种

不同类型的度分布的 GWD 统计值。这四种度分布类型是：(1)均匀分布；(2)线性增长分布；(3)线性递减分布；(4)指数递减分布。大多数的观测网络符合两类递减度分布中的一种情形。

在一个具有 55 个节点的网络中，由度分布类型及所选 α 值所带来的差异(图 2.6)，反映到 GWD 统计的取值上则显示为从 59.6 至 126.7。由于具有更高中心度的节点被赋予了更高的权重，因此，那些具有最高中心度的节点(呈现线性下滑态势)的网络就具有最大的 GWD 统计结果，而那些具有最低中心度的节点(呈现指数递减态势)的网络就仅能获得最小的 GWD 统计结果。



资料来源：改编自 M. Morris(个人通讯，2011 年 3 月 17 日)。

图 2.6　基于度分布和 α 值的 55 节点的 GWD 网络统计结果

几何加权边共享伙伴(GWESP)。第二个统计项 GWESP 是用来获取网络中的传递性模式的。边共享伙伴本质上就

是三角形构建；参见图 2.3 中对 1-ESP、2-ESP、3-ESP 结构的解释。GWESP 解释了在观测网络中聚类所对应的传递性特征。聚类(clusters)是指一群节点，它们内部紧密连接，而外部则甚少联系；这些聚类由一些三角形和拥有多个共享伙伴的边所构成。因此，GWESP 项能够检验这些三角形成为多个共享伙伴的边的趋势。GWESP 的定义为：

$$v(y;\alpha)=e^{\alpha}\sum_{i=1}^{n-2}\{1-(1-e^{-\alpha})^{i}\}ESP_{i}(y) \quad [2.10]$$

在公式 2.10 中，$ESP_i(y)$代表具有 i 个共享伙伴的边的数量(参见图 2.3)。除此之外，公式 2.10 与公式 2.9 的其他部分是完全一致的。我们仍选用癌症组织网络来举例，表 2.2 显示了不同 ESP 分布下(如图 2.4 所示)对应的 GWESP 统计的计算结果。

表 2.2　对图 1.3 和图 2.4 中所示的癌症组织网络的 GWESP 计算结果

ESP_i	Frequency $ESP_i(y)$	$\{1-(1-e^{-\alpha})^i\}ESP_i(y)$		
		$\alpha=0.25$	$\alpha=0.5$	$\alpha=1.0$
0	23	0.00	0.00	0.00
1	23	17.91	13.95	8.46
2	20	19.02	16.90	12.01
3	11	10.88	10.33	8.22
4	8	7.98	7.81	6.72
5	4	4.00	3.96	3.60
6	5	5.00	4.98	4.68
7	5	5.00	4.99	4.80
$\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^i\}ESP_i(y)$		69.79	62.93	48.49
$e^{\alpha}\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^i\}ESP_i(y)$		89.62	103.75	131.81

和 GWD 统计一致，有两个因素会影响到 GWESP 的统

计结果:网络中具有高 ESP 值的节点的比例以及所选择的 α 的值。由于两种统计项(指 GWD 和 GWESP)在形式上极为相似,GWESP 对于具有 i 共享伙伴的边的展现模式与 GWD 展现具有 i 中心度的节点的方式相同。所以,在一个拥有 55 个边共享伙伴的网络中,GWESP 的取值将会依据 ESP 分布以及所选择的加权参数(参见图 2.7)而不断增加。即那些包含更多多元共享伙伴(multiple partners)的边的网络将会具有更高的 GWESP,同时,α 值越高也会使得 GWESP 具有更高的值,尤其是在网络中多条边都具有多元共享伙伴时。

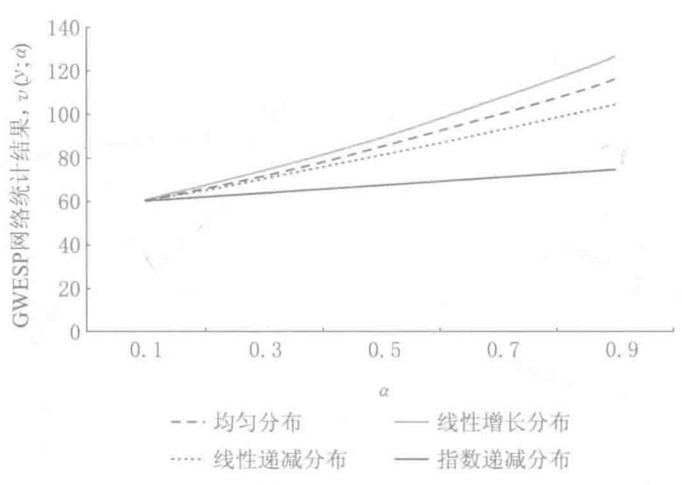


图 2.7 基于 ESP 分布和 α 值的 55 节点的 GWESP 网络统计结果

几何加权二元组共享伙伴(GWDSP)。最后,GWDSP 统计项关注的是具有共享伙伴关系的二元组数量(参见图 2.3)。这里,DSP 也包含在这些聚类中。GWDSP 术语被定义为:

$$w(y;\alpha)=e^{\alpha}\sum_{i=1}^{n-2}\{1-(1-e^{-\alpha})^{i}\}DSP_{i}(y) \quad [2.11]$$

这里 $DSP_i(y)$ 代表了那些与 i 具有共享伙伴(邻居关系)的二元组的数量(Hunter, 2007),公式 2.11 的剩余部分与公式 2.9 以及公式 2.10 是一致的。实际中,通常是那些具有多个二元组,且二元组都包含多个共享伙伴的网络会具有更大的 GWDSP 值;类似地,一个更高的 α 值也会赋予那些具有多个共享伙伴的二元组以更高的权重,最终导致 GWDSP 值的提升。如表 2.3 所示图 2.4 的癌症组织网络的 DSP 分布就体现了上述特征。同样的情况也可以在表 2.1 和表 2.2 中观察到。

表 2.3 对图 1.3 和图 2.4 中所显示的癌症组织网络的 GWDSP 计算结果

DSP_i	$Frequency\ DSP_i(y)$	$\{1-(1-e^{-\alpha})^i\}DSP_i(y)$		
		$\alpha=0.25$	$\alpha=0.5$	$\alpha=1.0$
0	356	0.00	0.00	0.00
1	167	130.06	101.29	61.44
2	94	89.40	79.45	56.44
3	31	30.66	29.11	23.17
4	29	28.93	28.30	24.37
5	13	12.99	12.88	11.69
6	6	6.00	5.98	5.62
7	7	7.00	6.99	6.72
$\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^i\}DSP_i(y)$		305.05	264.00	189.44
$e^{\alpha}\sum_{i=1}^{n-1}\{1-(1-e^{-\alpha})^i\}DSP_i(y)$		391.69	435.26	514.95

在上述三种几何统计项中,参数 α 的确定通常是采用事先指定的或者在建模过程中估计的方式获得的。通过后一种方式所建立的模型,被称为曲线指数族(curved exponential family, CEF)模型(Hunter, 2007)。如果采用事先指定方法获取 α,学者们通常会推荐分别尝试 $\alpha=0.25$ 和 $\alpha=0.75$,这

往往可以获得比较理想的结果。而一个更规范的建议是从 $\alpha=0.1$ 开始,"然后逐步尝试增加 α 的值,直到模型的对数似然估计值不再增长"(Goodreau, Handcock, Hunter, Butts & Morris, 2008:17)。在这一过程中需要记住的是:就上述三种几何统计项而言,α 值越大,网络统计值就越大。

第3节 | 本章小结

本章从历史的视角考察了 ERGM 的发展历程。ERGM 模型发轫于 1959 年的简单随机图模型，历经 50 年的发展，现在的 p^* 模型已经能够将高序依赖性条件纳入其分析框架中。ERGM 术语是在简单随机图模型和 p_1 模型提出之后被采纳的，但由于 ERGM 家族具有良好扩展性，能够包含之前已经形成的这些模型。因此，在 ERGM 发展阶段的中期，学者们采纳了 ERGM 这个术语来界定 ERGM 所属的广义统计家族，实际上，公式 2.6 和公式 2.7 中所表示的模型形式就表征了整个 ERGM 家族。下一章将着重描述一个复杂 ERGM 的建模过程。

第3章

建立一个有效的指数随机图模型

基本上，所有的模型都是错误的，但其中有些是有效的。

——博克斯和德雷珀(Box & Draper，1979：424)，

引自博克斯和德雷珀(Box & Draper，2007)

数十年来，网络科学家一直致力于改变一种现状，即现有的统计网络模型(如简单随机图模型)在解释真实社会网络的结构特征方面无法取得良好的效果。而马尔科夫依赖假设的应用及发展，可以帮助研究人员在统计网络模型建立的过程中，引入更为宽泛与复杂的依赖性假设，这一点对于研究人员展现、解释以及预测所观测的社会结构是十分有益的。虽然，指数随机图模型(ERGM)与基于二值数据的逻辑回归模型所依赖假设条件有所区别，但两者在模型的解释上的确具有较大的相似性。即，网络的连线被视为一种输出(不再被视为输入)，而网络的成员属性以及结构特征有助于解释、预测一条连线形成的概率(Hunter，Goodreau & Handcock，2008)。

接下来这部分将展示一个复杂指数随机图模型的构建过程。在模型构建之初，首先通过探索性分析，识别观测网

络的特征，并且获取在模型构建过程中具有重要影响意义的成员信息。指数随机图模型的构建过程是从简单随机图模型开始的，此时，简单随机图模型仅考察网络的密度指标；随后，通过添加主效应（main effects）和交互（interaction）统计项的方式将网络成员的属性特征纳入到模型中来，该步骤完成后将会形成一个二元独立性模型；最终，几何统计项将作为主效应和交互统计项的补充被纳入到模型中来，弥补前述模型在获取网络结构特征上的不足，从而形成一个新的依赖性模型。另外，在构建模型的过程中，本章也会穿插介绍针对模型拟合优度评价、模型诊断工具与策略以及模型结果的解释等内容。

本书的附录A部分（可以在线获取）包含了一个可用于复制分析过程的R命令列表，具有编号和标记的代码都可以从附录A中获取。本书中凡是标注了"Command 1"的地方，对应附录A中标注为"Command 1"的代码，利用该代码可以复制命令运行的结果。需要注意的是，由于本书中用于执行分析任务的软件是开源的，因此，这些软件并不是一成不变的。本书后续部分包含的对应软件及软件包的一些命令集可能需要根据软件版本的变化而进行调整。通过阅读相关命令的帮助文档，我们就能够了解这些命令变化的情况。在本章以及下一章节中存在少数的情况，书中段落部分可能会包含一些命令，这些出现在段落文字中的R命令是用courier字体来书写的；如果文字带有下划线，读者就需要用特指的文件名称或者其他信息替换这些文字。例如，命令"`read.paj('data file')`"就提示读者在使用该命令时，应采用一个数据文件的真实名称来替换data file这个词。

第1节 | 软件获取与准备

多种软件包都可以用于统计网络模型的估算，包括 PSPAR，Multinet，R-statnet，RSiena，以及 Pnet（Shumate & Palazzolo，2010）。下面的分析就是利用 R-statnet 包来实现的，R-statnet 是用 R 建立指数随机图模型的一个软件工具集，这个工具集的开发者人员列表可以在 statnet 的网页上查到（http://statnet.csde.washington.edu/about_us.shtml）。R 是一款免费的软件，它可以通过统计计算网站所提供的 R 项目获取，其网址是：http://www.r-project.org/。R 软件的定位是作为一个供开发人员使用的平台，开发人员能够轻松地在 R 架构的基础上开发并发布适合统计分析的软件包。用户除了需要安装 R 之外，还需要安装 statnet 套件，因为该套件是独立于 R 核心架构的。顾名思义，R-statnet 套件是由 statnet 团队所开发的，该套件包括 ergm，network，sna，以及 networkDynamic 等多个软件包。此外，R-statnet 套件还囊括了一系列为 statnet 套件提供支持的软件包，包括 robustbase，Martix，lattice，trust，nlme 以及 coda 包，这些包都是由 statnet 团队之外的人员（或团队）开发的，这其中每一个包都包含一些特殊的公式、函数以及有用的术语，这对于开发指数随机图模型而言是十分有益的。这些软件包的帮助文

档可以通过在 R 提示符之后输入 help(package)获得，注意需要将 package 转换为拟使用的软件包的名称。本文所展示的分析部分是在 R 第 2.15.2 版本下完成的，同时也使用了 statnet 套件的 3.01 版。

想要安装 R-statnet，可以通过在 R 下使用软件包安装菜单，或者在 R 提示符之后输入下列代码实现：

```
install.packages('statnet')        Command 1
```

该命令将会从存储 R 包的众多资源库中选择一个资源库作为安装 statnet 软件包的来源库。这些资源库也被称为“综合 R 存档网络”(Comprehensive R Archive Network，简称为 CRAN)，该网络分布在世界各地(http://cran.r-project.org/)。每一个 CRAN 站点都包含相同的资源，这些资源包括 R 软件包以及文档。如果已经安装过 statnet 套件，则可以用 update.statnet 命令对当前的 statnet 套件进行更新，update.statnet 命令也包含在“Command 1”(命令 1)中。

statnet 套件安装完毕之后，在每次开始使用 R 软件时，我们还需要导入 statnet 套件。因为，只有那些被导入到 R 内存的软件包才能运行。导入 statnet 套件的实现步骤，可以通过在 R 提示符后输入如下代码实现：

```
library('statnet')        Command 2
```

第2节 数据获取

本章的分析部分将会使用一个网络数据集，该数据集可以从美国国家城镇卫生官员协会(NACCHO)网站获取，该数据也可以从CRAN上通过下载R包的方式获得。下面的操作步骤与之前的操作步骤相似。首先，安装并打开"ergmharris"包进而获取数据；随后，利用 install.packages('ergmharris')命令安装数据；接下来利用 library('ergmharris')命令导入数据。在statnet套件和其他R软件包中还包含一些其他的数据集，这些数据集也会被用到；通过在R提示符后输入如下代码，可以观察当前的R列表中有哪些可以获得的数据集：

```
data( )                                    Command 3
```

在输入上述命令后，一个新的窗口会弹开显示R中可获得的所有数据集的列表，这个列表会根据所安装的数据包的差异有所变化。R能够支持将各种格式的网络数据导入。例如，在Pajek网络软件中保存的文件是以.paj或者.net为后缀的文件，这些文件就可以利用 read.paj( )函数导入。又如，在同一网络中，被保存为边列表形式的数据也可以在R中作为矩阵格式被导入。

当网络数据被导入R后，我们还需要根据导入数据的类

型，对相关文件进行转化，使其转化为特定网络数据类型，或者是在 R 语境下的网络数据类型（network class）。在 R 提示符之后输入 class(data name)就可以检查导入数据类型。如果反馈结果不是“network”，那么，在使用 statnet 套件进行网络模型构建之前，就需要将该数据转化为一个网络类型。数据转化方式依据数据格式的不同存在着差异，最简单的方法可以使用命令 as.network(data name)将数据转换为网络类型数据。将不同数据类型转化为网络对象的具体做法可以参见布茨（Butts，2008）的相关论著。

在这一章后续部分将会使用的 NACCHO 数据集，该数据集是一个针对全美地方卫生机构领导人之间的沟通关系网络的数据集。该数据是针对 2010 年 NACCHO 目录下全部的地方卫生机构（LHDs）进行问卷调查后获取的（http://www.naccho.org/about/LHD/），问卷调查的内容涉及了解地方卫生机构的组织结构、财务状况、领导体制、人员配备以及在地方层面都开展了哪些健康项目等问题。在 R 的提示符之后输入“Command 4”（命令 4）所示的命令，可以打开在 R 包里 ergmharris 数据集中的 LHD 网络数据。

```
data(lhds)                    Command 4
```

在数据导入之后，可以在提示符后键入这个网络对象的名称，“lhds”，以检查该数据是否被正确导入了。

```
lhds                          Command 5
```

“Command 5”（命令 5）输出的结果是对该网络对象（lhds）的描述性统计，包括网络规模（vertices＝1 283），是否为有向网络（directed＝FALSE），该网络有多少条边（n＝2 708），以及该网络的其他相关信息。接下来输出的是网络

成员的属性变量名称(即节点的属性名称),在本书中,节点属性是包含在网络对象中,作为其构成的一部分来使用。就本例而言,这个由 1 283 家地方卫生机构组成的网络存在五种节点属性(又可以称为网络成员属性):state(州)、nutrition(营养项目)、hivscreen(艾滋病筛查项目)、popmil(辖区人口)和 years(领导履职年限)。这些属性的含义如下:

state(州):地方卫生机构所属的州。

nutrition(营养项目):利用二值变量表示地方卫生机构是否开展了营养相关的项目,nutrition=Y 表示开展,而 nutrition=N 则表示未开展。

hivscreen(艾滋病筛查项目):利用二值变量表示地方卫生机构是否开展了艾滋病筛查项目,hivscreen=Y 表示开展,而 hivscreen=N 则表示未开展。

popmil(辖区人口):地方卫生机构所辖的人口(百万人)。

years(领导履职年限):地方卫生机构的现任领导的履职年限,该数据是一个分类变量,包括四类数值:履职年限为 1 至 2 年,则 years=0;履职年限为 3 至 5 年,则 years=1;履职年限为 6 至 10 年,则 years=2;最后,履职年限为 10 年以上,则 years=3。

与古德鲁和他的同事所做的一样(Goodreau et al., 2008),我们可以通过"Command 6"(命令 6)获得更加完整的网络信息概要以及网络成员属性特征信息(表 3.1)。表 3.1 所显示的网络信息的概要首先包括了对一般性网络信息的描述,如网络规模、密度、是否为有向网络(例如,directed=FALSE);在上述一般性网络信息之后,是对该网络对象所包含的五种属性特征的描述性统计。通过对这些属性特征的统计,结果显

示大多数地方卫生机构都执行艾滋病筛查项目(Y＝804；N＝461)以及营养项目(Y＝941；N＝326)；对 popmil(辖区人口)进行统计的结果显示：地方卫生机构所辖人口数量在 550 到 1 010 万之间，其中，密苏里州和俄亥俄州拥有的地方卫生机构数量最多，分别是 73 和 72 个。

表 3.1 R 输出的网络信息概要(部分)

```
Network attributes:
  vertices = 1283
  directed = FALSE
  hyper = FALSE
  loops = FALSE
  multiple = FALSE
  bipartite = FALSE
  title = lhds
  total edges = 2708
  missing edges = 0
  non-missing edges = 2708
  density = 0.00329279

Vertex attributes:

 hivscreen:
   character valued attribute
   attribute summary:
  N   Y
461  804

 nutrition:
   character valued attribute
   attribute summary:
  N   Y
326  941

 popmil:
   numeric valued attribute
   attribute summary:
    Min.   1st Qu.   Median     Mean   3rd Qu.      Max.
 0.00055  0.01722  0.04094  0.15860  0.12870  10.11000
```

续表

```
 state:
   character valued attribute
   attribute summary:
   the 10 most common values are:
MO  OH  MA  IL  KS  NJ  WI  NC  FL  MN
73  72  66  63  63  62  62  57  49  49
  vertex.names:
   character valued attribute
   1283 valid vertex names

 years:
   integer valued attribute
   1283 values

No edge attributes


Network edgelist matrix:
        [,1] [,2]
   [1,]   2   10
   [2,]   2   11
```

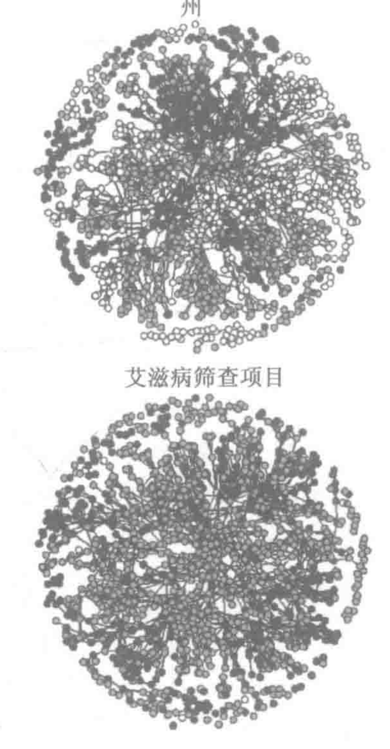
在网络信息概要表的节点属性特征之后是边属性特征统计。边属性统计是对网络中每条边所附加特征的统计。例如，如果网络数据中包括了两个地方卫生机构之间物理距离的信息，那么，一个边属性就可以用来标识出两个地方卫生机构之间相距的英里数。本例中，地方卫生机构的网络数据集中并不包括边属性。在表 3.1 中的最后一部分信息是网络的边列表信息。例如，边列表信息中的第一条边就是由节点 2 到节点 10 的连线所构成的，而该连线又表明了两个地方卫生机构之间的联系。表 3.1 中边列表信息由于长度的原因被截断了，但如果在 R 中通过运行“Command 6”则可以看到完整的结果。

第 3 节 | 数据探索

通常，在进行网络模型构建之前最好先进行数据探索工作。具体到网络数据，在网络模型构建和赋值的过程中，图形化展示以及描述统计等方法对于了解网络的结构特征是十分有帮助的。

复制附录 A 中标注"Command 7"(命令 7)的命令，会输出一幅网络图形，该图形通过对节点着色的方式显示地方卫生机构的属性特征，便于我们识别具有不同属性特征的地方卫生机构之间的关系模式。对州(state)特征进行着色的网络图中出现了明显的聚集现象，图中具有同样颜色的节点往往聚集在一起，说明那些处于同一区域的地方卫生机构往往会聚集在一起，这可能意味着一个地方卫生机构更愿意和与它处于同一州的地方卫生机构进行交流(图 3.1)。图 3.1 中对艾滋病筛查项目(hivscreen)进行着色的网络图同样也显示了一些聚集现象：色调较浅的地方卫生机构聚集在网络的中间区域，而色调较深的地方卫生机构则似乎处于网络的边缘。通过对这些图的分析，我们会产生一种假设：同一州内地方卫生机构建立沟通关系的概率要高于地方卫生机构之间随机建立沟通关系的概率。同样地，执行同类项目(如艾滋病筛查项目)的地方卫生机构的领导之间建立沟通关系的

概率要比领导之间随机建立沟通关系的概率高。需要注意的是，命令 7 所制作的网络图并不会产生和图 3.1 所示一模一样的节点的空间分布；该图利用点和连线来表示数据，但是每个节点的空间位置是随机的，并没有特殊的含义。

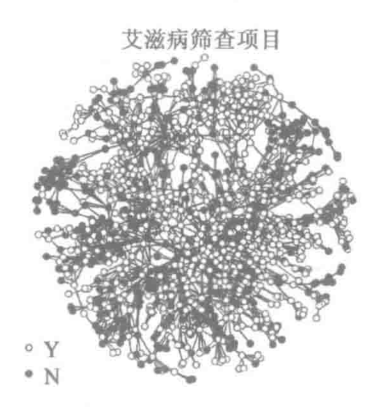


地方卫生机构网络描述了卫生机构之间沟通的情况；该网络图通过着色的节点来表示这些地方卫生机构的特征。

图 3.1

通常而言，网络中的节点数量过多往往会妨碍我们识别网络中的重要特征模式。因此，采用仅显示网络中最大成分

的方法(largest component,最大连通的节点集合)也许能够在一定程度上帮助我们厘清网络结构特征模式。利用“Command 8”(命令 8)可以筛选出网络的最大成分并将其绘制出来。这样一来,最大成分包含了网络中的绝大多数的节点(n=1 083),图 3.2 展示了网络中的最大成分,其中,网络中的节点根据地方卫生机构是否开展艾滋病筛查项目的结果被着色。注意图 3.2 增加了一个图例,该操作可以通过“Command 8”(命令 8)来实现;在 R 中,在提示符之后输入 help(legend),还可以选择很多其他参数来辅助建立和摆放图例。关于数据探索分析过程中的网络可视化展示的问题,以及其他的操作细节,可以从古德鲁与其同事(Goodreau et al., 2008)、布茨(Butts, 2008)以及 statnet 网站的介绍上获得(http://statnet.csde.washington.edu/)。



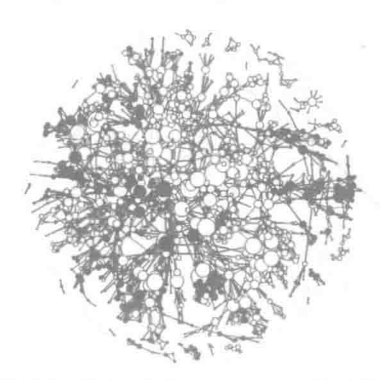
地方卫生机构网络中最大成分图,其中节点颜色根据地方卫生机构是否实施了艾滋病筛查项目判定。

图 3.2

节点的规模和形状是我们通过视觉来辨识网络特征模式

的其他方式。通常可以根据连续变量(continuous attributes)或者定序变量(ordinal attributes)的属性值来确定节点的大小;而根据定类变量(nominal variables)来确定节点的颜色及形状。在地方卫生机构网络中,数据属性类型主要为定类变量,而辖区人口以及领导履职年限这两项属性使节点依大小排列。另外,网络的测量结果,例如度也可以用来标识网络节点的大小。度是一名网络成员在网络中与其他成员建立联系的数量,因此,度就可以用来表示各个地方卫生机构沟通关系的数量。利用"Command 9"(命令 9)可以进行度属性测量,并利用该属性来绘制网络图。但遗憾的是,通过执行命令 9 将网络节点度属性值作为节点大小的绘制方法,会使得图中有些节点过大,且节点之间出现相互重叠现象,进而不利于对网络关系模式进行识别。

我们可以通过调整网络中心度测量结果来缩放节点大小。"Command 10"(命令 10)就采用了在原始中心度结果上除 6 的处理方式,于是,网络图中较大的节点表示那些与其他地方卫生机构有着更密切的联系的机构。网络图(如图 3.3 所示)已经初步展现出了与艾滋病筛查项目相关的网络关系模式,从中可以观察到:大的白色节点比大的黑色节点更多,这说明,开展艾滋病筛查项目的地方卫生机构(白色节点)与没有开展艾滋病筛查项目的地方卫生机构(黑色节点)相比,通常具有更高的网络中心度,即与其他地方卫生机构有更多的联系。对上述现象的一种可能解释是:那些开展艾滋病筛查项目的地方卫生机构往往处于大城市,因此,更易于与其他的地方卫生机构建立沟通关系。



地方卫生机构网络，其中，节点颜色根据是否执行艾滋病筛查项目判定，同时节点的大小依据地方卫生机构在整个网络中的联系数量(度)判定。

图 3.3

除图形化展示之外，对整体网络及其节点特征进行观察也可以为我们提供对于网络结构特征更深入的洞见，进而有助于我们确立网络的建模策略(Goodreau et al., 2008)。之前表 3.1 已经展示了网络的规模与密度，另外，还可以利用"Command 11"(命令 11)获得每个节点的平均连接数量(即度的平均数)、度值出现的频次，以及三元组的分布情况(参见图 1.5 中对于四种三元组类型的表述)。注意，这里对度的相关操作是以有向网络假定为前提的；如果网络是无向网络，则可以通过将"gmode"参数设定为"graph"的方式，指定图形为无向网络("digraph"表示有向网络)。

地方卫生机构网络中平均的度是 4.22(SD=2.90)。因此，一个地方卫生机构与平均 4.22 个其他地方卫生机构相互联系与沟通。通过观察我们发现，1 283 个地方卫生机构一共建立了 2 708 条沟通连线，平均连线数是 4.22 条，这可能比我们预想的要高。然而，一次单一的连接包括两个当地卫生机构，所以连接 A 机构与 B 机构之间的连线也会包含在 A 与 B

之间对中心度的计算中。实际上,2 708 条连线中的每条连线均会产生两次对地方卫生机构中心度的计数,因此,在由 1 283 个地方卫生机构构成的网络中,网络的度总值为 5 416(2 708 * 2)。总的度值 5 416 除以 1 283 个节点就得到地方卫生机构平均度值 4.22。度分布表的结果还显示:有 58 个地方卫生机构的连接次数为 0,有 117 家地方卫生机构仅与 1 家地方卫生机构建立了沟通关系等。另外,三元组统计表显示共有 347 709 795 个三元组完全没有连接关系,3 445 061 个三元组仅有一条连边,9 788 个三元组有 2 条连边,1 437 个具有完整的三角形。

```
> mean( degree( lhds, gmode = "graph" ) )
[1]  4.221356

> sd( degree( lhds, gmode = "graph" ) )
[1]  2.895897

> table( degree( lhds, gmode = "graph" ) )

  0    1    2    3    4    5    6   7   8   9  10
 58  117  182  223  226  172  104  67  35  25  26
 11   12   13   14   15   16   17  18  19
 14    8    6    8    4    3    1   1   1
 20   22
  1    1

> triad.census( lhds, mode = "graph" )
             0       1    2    3
[1,] 347709795 3445061 9788 1437
```

除上述基础统计之外,对度、边共享伙伴(ESP)以及二元组共享伙伴(DSP)的分布情况进行图形化观察,也有利于理解网络的结构特征。与大多数观测网络所呈现的度分布特征一致,地方卫生机构网络的度分布图也显示:观测网络中

包含大量的具有低中心度的节点和少量的具有高中心度的节点，当将地方卫生机构网络与具有同等网络规模与密度的随机网络进行比较时（利用 Command 12，结果参见图 3.4），我们发现两者存在较大差异。注意"Command 12"（命令 12）可能需要花费 10 分钟甚至更长的时间运行。

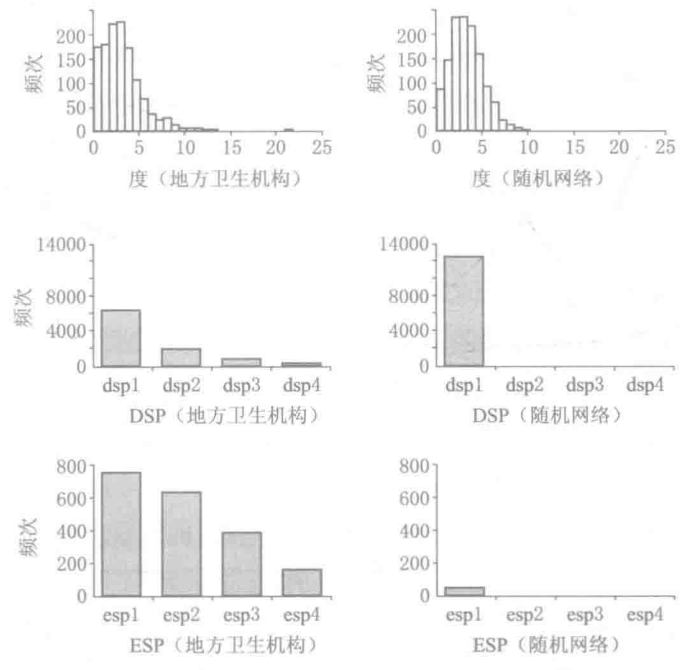


图 3.4 地方卫生机构网络(左)与具有同样网络规模和密度的随机网络(右)的度和共享伙伴分布图

另外，观测网络（地方卫生机构网络）与随机网络在边共享伙伴与二元组共享伙伴的分布上也表现出差异（参见图 2.3 中边共享伙伴与二元组共享伙伴的例子），在地方卫生机构网络中，更多的网络成员具有多个边共享伙伴与二元组共享伙伴，就这一点而言，随机网络与观测网络具有显著区别，

随机网络的特征是大量的节点仅具有单一的共享伙伴，而具有多个共享伙伴的节点几乎没有。

通过对网络进行图形化展现能够帮助研究者发现网络中潜在的聚集模式，而采用混合矩阵和相关系数方法是识别这种网络聚集模式的另一种方式。正如古德鲁及同事(2008)所论述的，混合矩阵可以针对一个分类属性变量各层次之间各种可能的组合形式进行统计，从而检验相互连接的二元组（例如两个地方卫生机构之间的联系）在连接属性上存在的特征。例如，满足"两个地方卫生机构均执行了艾滋病筛查项目"条件的二元组有多少？或者满足"一个地方卫生机构位于密苏里州而另一个位于加利福尼亚州"条件的二元组有多少？根据之前的图形化展示结果可知，当以州以及执行艾滋病筛查项目为依据对网络节点进行着色时，网络中存在一些潜在的网络聚集证据。混合矩阵可以帮我们确认这种网络聚集关系的模式，当然，也可以利用混合矩阵去探索其他的网络节点属性(Command 13，表 3.2)。

表 3.2 执行艾滋病筛查项目、营养项目以及领导履职年龄的混合矩阵

```
> mixingmatrix( lhds, "hivscreen" )
     N    Y
N  526  632
Y  632 1498
> mixingmatrix( lhds, "nutrition" )
     N    Y
N  216  648
Y  648 1812
> mixingmatrix( lhds, "years" )
     0    1    2    3
0   71  190  207  283
1  190  120  259  355
2  207  259  225  516
3  283  355  516  389
```

这些混合矩阵将网络中属性的层级作为矩阵的行和列。矩阵单元格中的数字表示矩阵中具有对应行和列属性的相互连接的二元组数量。例如，在混合矩阵中，两个均执行营养项目的地方卫生机构之间相互连接的数量是1 812，该数量显示在表3.2中第二个混合矩阵的右下角。另外，执行营养项目和没有执行营养项目的两家地方卫生机构之间建立沟通关系的次数是648次，这个数值记录在混合矩阵的对角。本文中没有展示针对州(state)属性的最全面的混合矩阵(共有49行乘以49列)，但其输出结果仍能够通过运行"Command 13"(命令13)从R的输出窗口中观察到。通过观察州属性(state)的混合矩阵我们可以发现：处于同一个州的地方卫生机构之间更易建立沟通关系。

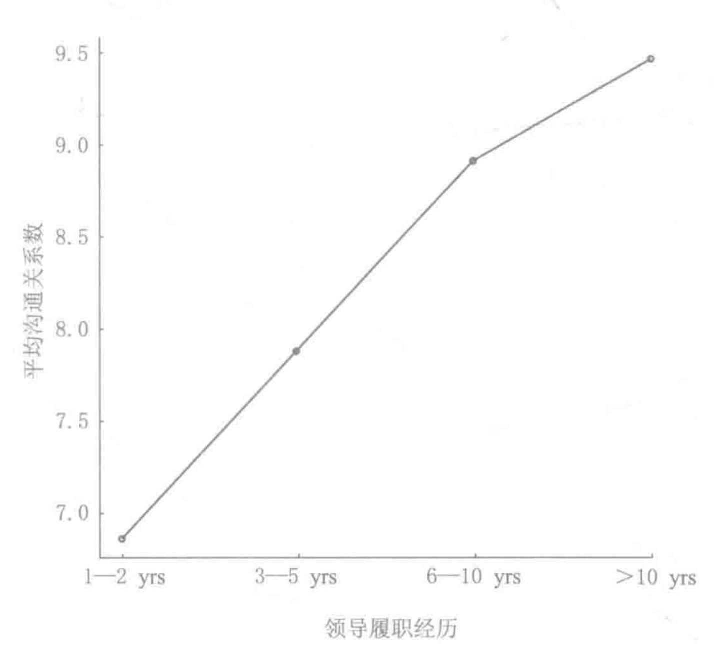
我们也观察到了在混合矩阵中出现的一些关系模式。在相连的二元组中，其中均执行了营养项目的两个地方卫生机构建立沟通关系所构成的二元组数量为1 812个，而一家执行了营养项目的机构和一家没有执行营养项目的机构之间建立沟通关系的二元组数量是648个。该现象揭示了网络的另一个特征：执行相同项目的地方卫生机构更有机会连接在一起(即，在执行项目方面存在的同质性)。艾滋病筛查项目的混合矩阵也显示了同样的模式，虽然没有营养项目那么显著，其中均执行了艾滋病筛查项目的两个地方卫生机构建立沟通关系的二元组数量为1 498个。然而，领导履职经验的混合矩阵则显示了一些差异化的连接模式，在领导履职经验混合矩阵中，那些更具经验的地方卫生机构领导似乎与所有履职年龄段的其他领导保持了较紧密的联系。例如，在已经形成沟通关系的地方卫生机构二元组中，包含"履职年

限为1至2年”经验领导(编码为0)的数量为751对,包含“履职年限为10年以上”的领导(编码为3)的数量283对(38%)。相较而言,有1 543对连通二元组包含“履职年限为6至10年”的领导,而仅有18.3%的连通二元组包含“履职年限为3至5年”的领导。这样,通过计算不同履职经验类型之间连通的平均数,并结合作图就可以进一步去检验领导履职经验与网络结构之间关联关系(Command 14;图3.5);图3.5显示,履职经验丰富的领导所在的地方卫生机构之间更易于建立沟通关系。

进一步对网络节点的特征进行观察可以为我们提供对网络结构更深的洞见。例如对连续型变量进行观察,如popmil(辖区人口/百万人),一种有效的方法是检验该变量与中心度之间的相关性(Command 15)。相关系数的结果为0.27。该结果显示:地方卫生机构所在辖区的人口数量越多,该机构与其他机构之间建立的联系就越多。另外,在一些网络中,也可能通过不同属性特征节点的平均连接数量来观察网络的结构特征。“Command 16”(命令16)利用双向表(two-way tables)来探索网络的属性数据。

上述用于识别地方卫生机构特征的探索性分析,对于模型构建是十分重要的。针对地方卫生机构网络的探索性分析结果显示,首先,机构在地域分布上存在较为广泛的同质性特征;而机构在项目执行方面则存在中等程度的同质性特征。其次,具有更多丰富履职经验的领导者的机构往往与其他机构建立了更丰富的联系;同样地,所在辖区人口数量越多的机构往往与其他机构建立了更丰富的联系。最终,地方卫生机构网络的基本结构特征显著不同于具有同等规模和

密度的随机网络。尤其表现在:度分布并不是均衡的,大部分节点仅有较低的中心度,而少数节点具有较高的中心度。同样,在地方卫生机构网络中,更多的网络成员具有多个边共享伙伴与多个二元组共享伙伴,这一点与随机网络也有着较大区别,这显示了传递性和前传递性(pretransitivity)特征在观测网络中要比随机网络中更为明显。在地方卫生机构网络中,同质性、非均匀的度分布以及传递性等都与现有的网络理论和模型构建策略是一致的。而当所有的这些特征都被融合进一个网络模型之下时,就能够更好地理解隐藏在真实观测网络结构之下的社会力量。



领导者的履职经验越多,地方卫生机构建立的沟通关系就越多。

图 3.5

第 4 节 | 模型构建

零模型

与其他模型的构建过程相仿，统计网络模型的构建也是以零模型（null model）为起点的。零模型是一个简单随机图模型，第 2 章描述了这个最简单的模型，该模型仅由一个单一统计项——网络的边或者是连线的数量——构成（Goodreau et al.，2008）。公式 2.6 中一般的 ERGM 模型公式可稍作修改用来表述零模型：

$$\text{logit}(P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c})) = \theta_{\text{edges}}\delta_{\text{edges}} \qquad [3.1]$$

其中，δ_{edges}表示以边数为变量的变化统计，而 θ_{edges}则表示边数统计项的系数。我们可以针对地方卫生机构网络构建零模型，采用“Command 17”（命令 17），其结果包含边数统计项的系数（$\theta_{\text{edges}}=-5.71272$）以及其他一些信息（参见表 3.3）。

利用公式 2.7，我们可以根据表 3.3 中所提供的信息计算出地方卫生机构网络中的任意一条连线（表示机构之间建立沟通关系）形成的概率。该模型仅考虑了一个条件，即网络的边数。标注为估计值（estimate）的列是专门存放模型的统计项所对应系数（θ）的地方。本例中，边数统计项的系数是

表 3.3　地方卫生机构网络的零模型

```
==========================
Summary of model fit
==========================

Formula:    lhds ~ edges

Iterations:  20

Monte Carlo MLE Results:
         Estimate Std. Error MCMC % p-value
edges  -5.71272   0.01925     NA  <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual  Deviance:   36365  on 822402  degrees of freedom
           Deviance: 1103728  on      1  degrees of freedom

AIC: 36367    BIC: 36379
```

负值(−5.71272),显示网络的密度是在50%以下,如果边数项的系数为0则表示网络具有50%或者0.5的密度。边数统计项的系数为负值是真实观测网络的典型特征,很少有观测网络具有0.5或者更高的密度,大多数网络模型的边数项的系数都为负值。

需要记住的是,变化统计(δ)代表当网络增加一条边时(即 Y_{ij} 从0到1变化时)相关统计项的变化情况(Hunter, Goodreau & Handcock, 2008)。网络的边数统计项(edges term)通常具有相同的变化统计值,$\delta_{edges}=1$。因为边数统计项是对网络中边数的考察,当网络增加一条边时,网络中的边数变化即为1。于是,我们就可以像计算逻辑回归模型一样,计算模型右侧的逻辑函数:$\frac{1}{1+e^{-(\theta_i X_i)}}$(Field, 2009)。

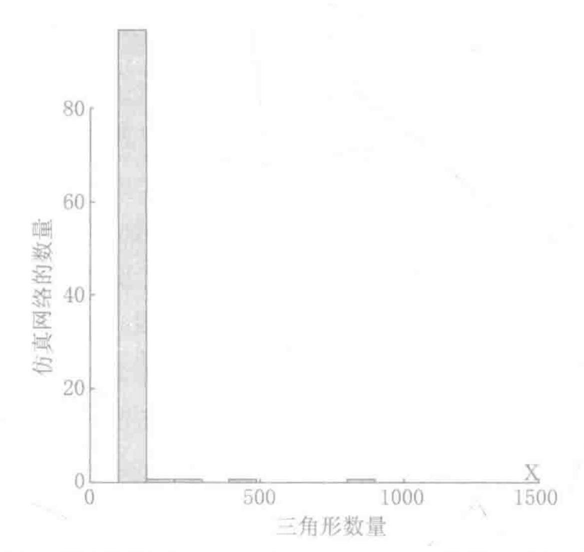
$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c) = \text{logistic}(\theta_{\text{edges}}\delta_{\text{edges}})$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c) = \text{logistic}(-5.71272 * 1)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c) = \frac{1}{1+e^{-(-5.71272*1)}} = 0.003293$$

正如预期的那样，一条边形成的概率与地方卫生机构网络的密度是保持一致的，0.0033。该模型利用了标准二值逻辑回归模型（standard binary logistic regression）采用的最大似然估计（maximum likelihood estimation）方法。由于零模型是一个简单随机图模型，因此，该模型没有考虑复杂的依赖性假设条件。虽然零模型方法看上去是用一种复杂的方法来描述网络密度这个最简单的网络特征，但零模型的构建为未来更为复杂模型的构建提供了拟合优度评价的基准，所以，构建零模型也是十分有价值的。

虽然零模型能够很好地表征所观测的地方卫生机构网络的密度特征、但零模型并不能很好表征观测网络的其他特征。以零模型为基础进行仿真网络的网络测绘图，有助于我们了解所构建的模型在表征网络结构特征（例如三角形）方面哪里做得比较好，哪里做得不好。依照古德鲁及其同事（2008）的思路，我们可以利用"Command 18"（命令 18），以零模型为基础进行仿真，构造 100 个随机网络，并绘制这些仿真网络所包含三角形数量分布情况。图 3.6 显示了根据零模型仿真所产生的 100 个网络的三角形数量分布。X 标记的是观测网络（地方卫生机构网络）包含了 1 437 个三角形，这个数量要远高于依据零模型仿真产生的 100 个随机网络中任意一个网络的三角形数量。很明显，这个网络中的传递性特征需要采用更复杂的模型来获得。



依据零模型仿真产生的100个随机网络的三角形数量。X则标注了地方卫生机构网络所包含的三角形数量。

图 3.6

增加节点属性

研究人员在改进拟合优度的过程中，首要考虑的因素是：节点的属性是否会影响到网络中连线形成的概率。放到我们现在所讲的例子，地方卫生机构特征和机构履职领导的特征是否会影响到地方卫生机构沟通关系的形成呢？描述性统计结果显示领导履职经验以及辖区人口数量可能会影响一个地方卫生机构沟通关系产生的数量。为了检验这些节点属性对于关系形成可能性的影响，我们将这些属性因素增加到模型中来，形成主效应(main effects)模型。针对地方卫生机构属性对网络沟通关系产生影响的主效应的假设检验条件可以采取如下方式表示：

H_0:地方卫生机构所辖区域人口与地方卫生机构之间建立沟通关系的可能性之间不存在关联。

H_1:地方卫生机构所辖区域人口与地方卫生机构之间建立沟通关系的可能性之间存在关联。

在增加主效应统计项的过程中，根据数据类型的差异选择适当处理命令十分重要。莫里斯、汉考特与亨特(Morris, Handcock & Hunter, 2008)提供了一个综合列表，该列表包括了 statnet 套件中可获得的统计项以及这些统计项的具体使用指南。就地方卫生机构网络而言，领导者的履职年限将被作为一个分类变量包含到模型中来，而辖区人口则被作为一个连续变量被包含到模型中来(Command 19)。

在 statnet 包中，分类型的主效应统计项可以通过 nodefactor* 纳入到模型中来，而连续型的主效应统计项则可以用 nodecov* 纳入到模型中来。选择 nodefactor 参数会为模型增添多个统计量，其中的每一个统计量分别代表具有某种专门属性的一个节点在边的任意一端出现的次数；而选择 nodecov 参数则仅为模型增添一个统计量，该统计量是对构成连线的两个节点相关属性值的求和。例如，边 ij 表示一个具有 120 万人口的地方卫生机构和一个具有 50 万人口的地方卫生机构之间建立了沟通关系，那么，随着这条边的增加而来的是，在辖区人口数的统计结果上，出现 120 万＋50 万＝170 万的人口变化。

* statnet 中的参数。——译者注

表 3.4 地方卫生机构的主效应模型

```
==========================
Summary of model fit
==========================

Formula:  lhds ~ edges + nodecov("popmil") + nodefactor("years")

Iterations:  20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC % p-value
edges               -6.22545   0.06353     NA < 1e-04 ***
nodecov.popmil       0.19663   0.01431     NA < 1e-04 ***
nodefactor.years.1   0.14379   0.04509     NA 0.00143  **
nodefactor.years.2   0.27927   0.04216     NA < 1e-04 ***
nodefactor.years.3   0.33689   0.03983     NA < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual  Deviance:   36166  on 822398  degrees of freedom
           Deviance: 1103927  on      5  degrees of freedom

AIC: 36176    BIC: 36234
```

要想解释表 3.4 的结果，首先需要了解网络中任意一个分类变量所对应参照组的含义。在 statnet 包中，一个指数随机图模型的默认参照组是网络统计摘要列表中所显示的第一组（参见表 3.1）。在本例中，"履职年限为 1 至 2 年"就是领导者履职年限的参照组*。与逻辑回归模型相似，主效应模型往往会省略这些参照组而直接计算出其他类型分类的估计结果。可以通过调整基本参数方式来改变参照组，例如，我们可以观察到，在"Command 19"（命令 19）中针对领导履

* 可参考前文"years 领导履职年限"变量的定义。——译者注

职年限统计项所设定的参数是 nodefactor('years'),意味着该统计项中除了属性名称之外没有包含任何其他参数,因此,默认的分类变量中第一类就会被自动划为参照组。如果想选择最后一个分类,即,"履职年限为 10 年以上",那么就需要增加一个值为 4 的基本参数,以显示该分类属性变量的第四类将会作为参照组,如下所示:nodefactor('years', base=4),我们可以使用"Command 20"(命令 20)运行模型和总结命令,就像前面所示的那样。统计摘要表将包含领导履职年限的前三类,而省略具有最高领导履职年限的分类作为其参照组。需要注意的是,表 3.4 中包含四项在统计学意义上具有显著性的主效应项(辖区人口项及三类领导履职年限项)。

使用 R 软件进行工作的一个好处是:用户可以直接修改 R 软件底层的代码,而这些代码是由软件包的开发人员所提供的。这样一来,用户能很方便地设定统计概要表中所需要呈现的报告选项。例如,在某一领域中,如果统计检验对于报告逻辑回归模型是一个标准步骤,那么,用户就可以修订指数随机图模型的统计概要函数,从而实现当统计摘要函数运行时,就能得到每一个系数的统计检验结果,而不用通过额外的命令来产生这些统计检验结果信息。修改指数随机图模型统计摘要函数的 R 代码,可以使用"fix( )"命令(参见 Appendix B,在线版)。

根据附录 B 第一部分所提供的技术指导,我们可以修订指数随机图模型统计摘要函数的代码,并且仅仅运行命令 19,从而获得一个主效应模型的统计摘要表,其中包含了沃尔德检验的统计结果。在此基础上,我们可以对统计摘要函

数进行编辑从而改变模型摘要表中所显示的内容，而无须对模型进行重新估计。需要提醒的是，如果开发人员在未来的 statnet 包的版本中修订了相关环境，那么，通过“`fix( )`”来修改底层代码的命令也有可能会改变。于是，作为获取统计检验结果一种替代方法，我们也可以采用“Command 21”(命令21)的后半部分所采取的方法(提醒：该代码只有在命令 21 前半部分代码已经运行的情况下才有效)。

在主效应模型中(参见表 3.4)，所有的参数估计结果都是显著且正向的。此结果意味着：当领导的履职经验更为丰富时，地方卫生机构之间建立沟通关系的可能性就会增加；同样地，如果地方卫生机构所在地域拥有更多的辖区人口时，地方卫生机构之间建立沟通关系的可能性也会增大。这些结果和之前采用混合矩阵和相关系数方法对辖区人口及领导履职年限进行判断的结果是一致的。

在解释系数的含义方面，除了可以利用显著性及系数正负向这样的一般性解释方法之外，模型系数和它们对应的标准差(standard errors)也可以被转化为优势比(odds ratios)和置信区间(confidence intervals)，作为每个属性所对应系数的解释方法(参见 Command 21 以及表 3.5)。为了实现系数的转化，我们只需简单使用指数转化的方法(e^{θ})。通常，优势比会伴随着置信区间出现，主要用于描述模型估计的显著性和精度。参数 95%的置信区间可以根据如下方式计算：

$$95\%\mathrm{CI}_{\theta} = e^{\theta \pm 1.96 s.e._{\theta}} \qquad [3.2]$$

对于略大或略小些的置信区间(例如，99%的置信区间或 90%的置信区间)，用适当的 z 值取代 1.96，z 的取值分别

是2.56和1.28。优势比需要根据分类变量的参照组来进行解释。对于连续变量而言，优势比被定义为：相关变量每增加一个单位时，模型统计结果出现概率的变化情况。如果优势比大于1表明结果出现概率的增加；当优势比小于1时则表明结果出现概率的减小；当优势比等于1则显示变量与结果之间没有联系。因此，当置信区间包含1时也显示了变量与结果之间并不存在显著的关联。非显著的优势比以及边数统计项的优势比都可以在统计摘要表中展现，但这些内容一般不作解释。

表3.5 主效应模型参数的优势比及95%的置信区间

	Lower	OR	Upper
edges	0.0017	0.0020	0.0022
nodecov.popmil	1.1836	1.2173	1.2519
nodefactor.years.1	1.0570	1.1546	1.2613
nodefactor.years.2	1.2173	1.3222	1.4361
nodefactor.years.3	1.2954	1.4006	1.5143

根据主效应模型，当其他网络特征保持不变时，具有3到5年履职年限的领导所处的地方卫生机构与某个地方卫生机构建立沟通关系的概率，是仅具有1到2年履职年限的领导所处机构与之建立沟通关系概率的1.15倍。其中，参数95%的置信区间的范围是1.06到1.26，表明机构之间关系的真实值可能处于这一范围。同样地，当其他条件不变时，拥有超过10年履职年限领导的地方卫生机构与某个地方卫生机构建立沟通关系的概率，是仅拥有1至2年履职年限的领导所处的机构与之建立沟通关系的概率的1.4倍。

除了根据系数和标准差计算优势比以及置信区间对模型进行估计之外，R-ergm程序也提供了许多可以用来解释

与展现模型特征的其他对象。通过输入对象名称的方式"Command 22"(命令22),可以获得R-ergm模型中所包括的对象的列表。R-ergm的帮助文档对所有在描述列表出现的具体对象都有描述。

在大多数情况下,对逻辑回归模型进行估计需要报告优势比(ORs)的相关信息。因此,有必要将优势比以及置信区间所对应的列值嵌入到默认统计概要中去。附录B的第二部分提供了修改ERGM统计摘要表的技术指南,使统计摘要表能够包括优势比和置信区间信息。表3.6展现的正是表3.4的扩展版本。

伴随着模型统计摘要表的输出结果,我们就能够检验最初假设:

H_0:地方卫生机构所辖区域人口与地方卫生机构之间建立沟通关系的可能性之间不存在关联。

H_1:地方卫生机构所辖区域人口与地方卫生机构之间建立沟通关系的可能性之间存在关联。

根据主效应模型的统计结果,拒绝H_0假设而支持H_1假设($p<0.05$),地方卫生机构所辖区域人口与地方卫生机构之间建立沟通关系的可能性之间存在显著关联。具体而言,假定其他条件不变的情况下,辖区内人口数量每增加100万,该辖区所在的地方卫生机构形成沟通关系的概率将会增加1.22倍(OR=1.22; 95%CI=1.18−1.25)。

模型的概率预测

和零模型类似,主效应模型也可以用于预测任何两个网

络成员之间关系形成的概率。由于之前构建的模型已经将网络成员的属性纳入进来，因此，模型就可以计算具有某种属性特征的网络成员之间建立联系的概率。对于主效应模型的自变量(predictors)而言，每一个统计项的变化统计结果是较为直观的，如果自变量是分类变量时，那么，变化统计的结果是0、1或者2。如果二元组中两个网络成员均不具有相关属性特征，那么，变化统计值是0；如果二元组中仅有一个网络成员具有相关属性特征，那么，变化统计值是1；如果二元组中两个网络成员均具有相关属性特征，那么，变化统计值是2。于是，在地方卫生机构网络中，两个具有丰富领导履职经验的机构之间建立沟通关系的概率等于nodefactor.years.3的系数乘以变化统计值2的结果，其中，nodefactor.years.3表示具有丰富领导履职经验的机构，而变化统计值2则表示两个机构均具有相关的属性特征。而一个具有丰富领导履职经验的机构与一个刚履职领导所在机构建立沟通关系的概率，就等于nodefactor.years.3乘以变化统计值1，以此类推。依照亨特、古德鲁及其同事(2008)所提出的标识规则，δ符号对应一个分类型的点属性特征变量，具体可以表示为：

$$\delta_{cat}=\begin{cases}2,\text{节点 } i \text{ 与 } j \text{ 均具有该属性特征}\\1,\text{节点 } i \text{ 或 } j \text{ 具有该属性特征}\\0,\text{节点 } i \text{ 与 } j \text{ 均不具有该属性特征}\end{cases}$$

如果自变量是连续型变量，δ则表示二元组中两个地方卫生机构领导均具有特征数的和。就地方卫生机构网络而言，辖区人口数量就是一个连续型的自变量。所以，当一个

地方卫生机构所在地域拥有 100 万人口，而另一个机构（所在地域）仅拥有 50 万人口时，那么，人口数量所对应的 δ 参数（δ_{popmil}）就是 1+0.5=1.5。

为了预测两个地方卫生机构之间建立沟通关系的概率，例如，机构(1)所在的辖区拥有 200 万人口（popmil=2），且机构的领导具有 7 年履职年限（years=2）；机构(2)所在的辖区拥有 10 万人口（popmil=0.1），且机构的领导仅有 1 年的履职年限（years=0）。表 3.6（统计摘要表）中估计列包含系数需要乘以对应的各统计项的变化统计值。

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c})$$
$$=\text{logistic}\begin{pmatrix}\theta_{edges}\delta_{edges}+\theta_{popmil}\delta_{popmil}+\theta_{3-5years}\\ \delta_{3-5years}+\theta_{6-10years}\delta_{6-10years}+\theta_{>10years}\delta_{>10years}\end{pmatrix}$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c})$$
$$=\text{logistic}(-6.23*\delta_{edges}+0.20*\delta_{popmil}+0.34*\delta_{6-10years})$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c})$$
$$=\text{logistic}(-6.23*1+0.20*2.1+0.34*1)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c})$$
$$=\text{logistic}(-6.08)=\frac{1}{1+e^{-(-6.08)}}=0.0023$$

两个具有上述特征的地方卫生机构之间建立沟通关系的概率为 0.0023 或 0.23%（参见图 3.7）。虽然，这个概率看起来很低，但需要注意的是，该网络的密度是 0.0033，意味着原先预测的地方卫生机构网络建立沟通关系的概率事实上有 1/3 个 1%。因此，这里所描述的两个地方卫生机构之间建立沟通关系的概率要比机构间建立沟通关系预计的概率低。

表 3.6 展开的主效应模型统计摘要表

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years")

Iterations:  20

Monte Carlo MLE Results:
                    Estimate Std. Error MCMC %    Lower       OR Upper p-value
edges               -6.225446 0.063528      NA 0.001747 0.001978 0.002 < 1e-04 ***
nodecov.popmil       0.196630 0.014310      NA 1.183626 1.217293 1.252 < 1e-04 ***
nodefactor.years.1   0.143793 0.045086      NA 1.056988 1.154645 1.261 0.00143  **
nodefactor.years.2   0.279269 0.042164      NA 1.217290 1.322163 1.436 < 1e-04 ***
nodefactor.years.3   0.336894 0.039827      NA 1.295417 1.400590 1.514 < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   36166  on 822398  degrees of freedom
          Deviance: 1103927  on      5  degrees of freedom

AIC: 36176    BIC: 36234
```

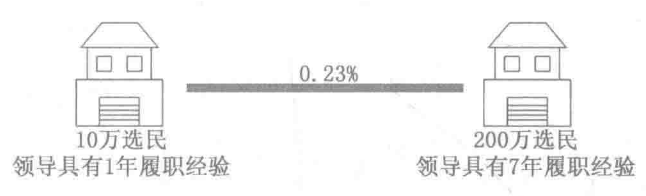


图 3.7 基于主效应模型两个地方卫生机构之间建立沟通关系的概率

增加交互项

节点属性说明的是每一个网络成员的个体特征，而针对节点属性的交互项则关注一个二元组中两个网络成员的属性特征(Morris et al., 2008)。最常用的交互项主要关注的是：二元组中两个网络成员间的同质性问题(两个节点均具有的属性，例如都是男性)或者异质性问题(两个节点具有不同的属性，例如一个网络成员为男性，另一个为女性)。

交互项会将一个二元组视为一个独立组成部分，因此，包含了交互项的统计网络模型就成为了二元独立性模型。需要记住的一点是：二元独立性模型假定网络中的每一个二元组都是独立于模型中的其他二元组的，所以，帕姆和米歇尔之间具有联系的概率是独立于菲尔和帕姆之间具有联系的概率的，即使帕姆同时存在于两个二元组中。

基于此前的探索性分析，如果两个地方卫生机构同属于一个辖区，并且都执行同样的项目，那么，这两个地方卫生机构之间建立沟通关系的概率似乎更高一些。也就是说，在地方卫生机构的网络中，似乎存在机构之间基于地域和项目执行的同质性倾向。在新模型(二元独立性模型)中我们针对这些属性特征采用交互项的方法来检验同质性倾向的假设

(Command 23)。

模型的结果显示:地方卫生机构网络在地域和项目执行方面的系数为显著且正向的。也就是说,两个位于同一区域的地方卫生机构更有可能建立沟通关系;同样,两个执行同一项目的机构之间也更有可能建立沟通关系。在R命令中设定"nodematch"参数的方式可以将同质性统计项的结果导出,结果如表3.7中高亮的部分所示。统计摘要表中所有的主效应项(包括nodefactor项以及nodecov项)均显示为正向且显著。需要注意的是,执行同一项目的主效应统计项并没有被纳入到该模型中来,因为一个地方卫生机构是否执行一个项目仅可以获得两种可能的结果(Y, N),鉴于目前网络有限的自由度,我们不可能对同一数据既进行主效应项测量又进行交互项的测量。还有一些其他的统计项也可以支持进行额外主效应和交互效应检验(Goodreau et al., 2009; Morris et al., 2008)。

根据探索性分析,我们可以得出:执行同一项目的地方卫生机构之间具有的同质性与不执行项目机构之间具有的同质性是有差异的。地方卫生机构更可能与那些执行了同样项目的机构建立沟通关系,但反之则未必如此。也就是说,没有执行某一项目的机构未必会和那些没有执行某一项目的机构之间建立沟通关系。因此,对于同质性的估计可以在分类变量所包含的类别层次上进行,这就是所谓的差异化同质性(differential homophily)。通过对机构的项目进行差异化同质性说明,同质性项将会被区分为执行项目和不执行项目,可参见"Command 24"(命令24)和表3.8。差异化同质性估计的实现方法是在nodematch对应属性名称的后面加

表 3.7 包含同质性统计项的地方卫生机构网络模型

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years") + nodematch("hivscreen") +
    nodematch("nutrition") + nodematch("state")

Iterations:  20

Monte Carlo MLE Results:
                      Estimate Std. Error MCMC %     Lower        OR   Upper  p-value
edges               -9.537e+00 1.133e-01    NA 5.775e-05 7.212e-05   0.000  < 1e-04 ***
nodecov.popmil       3.643e-01 1.939e-02    NA 1.386e+00 1.440e+00   1.495  < 1e-04 ***
nodefactor.years.1   1.810e-01 4.743e-02    NA 1.092e+00 1.198e+00   1.315  0.000135 ***
nodefactor.years.2   3.249e-01 4.435e-02    NA 1.269e+00 1.384e+00   1.510  < 1e-04 ***
nodefactor.years.3   3.040e-01 4.207e-02    NA 1.248e+00 1.355e+00   1.472  < 1e-04 ***
nodematch.hivscreen  2.889e-01 4.668e-02    NA 1.218e+00 1.335e+00   1.463  < 1e-04 ***
nodematch.nutrition  2.754e-01 4.665e-02    NA 1.202e+00 1.317e+00   1.443  < 1e-04 ***
nodematch.state      6.279e+00 8.491e-02    NA 4.514e+02 5.332e+02 629.736  < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   19568  on 822395  degrees of freedom
          Deviance: 1120525  on      8  degrees of freedom

AIC: 19584    BIC: 19677
```

上 diff=T。

如果自项目分析之初到目前为止，你还没有关闭或重启过 R 程序的话，那么，你目前的模型估计可能已经运行得十分缓慢了，或者已经出现了内存分配错误的情形。针对这种情况，有一些策略可以帮助你在不关闭或重启 R 的情况下提升 R 的运行速度。首先，你可以移除不再需要的对象，为了实现这一步骤，你首先需要使用 ls( )命令来列出当前所有 R 中打开的对象，从该清单中识别出你已经不再需要调用的对象，并使用命令 remove(object)逐一删除它们。一旦删除了这些不再需要的对象，你就可以运行 gc( )命令通过垃圾回收站来清理 R 的内存。最终，在对象删除完毕而且清理工作已经完成时，你就可以使用 memory.size(8000)命令增加 R 内存的分配。

需要注意的是，表 3.8 中已经高亮显示出了每一类同质性统计项的统计结果。包含了差异化同质性的模型结果显示：两个执行了同一项目的机构之间建立沟通关系的概率呈现了显著的增加态势，而两个没有执行同一项目的机构之间建立沟通关系的概率则不存在显著性特征。

在某些情况下，我们可能仅需要保留执行了同一项目的机构之间的同质性统计项，而不必保留那些没有执行同一项目的机构之间的同质性统计项。这可以通过明确你想要保留在 nodematch 命令中的项来实现。这种情况下，我们将那些没有执行同一项目的机构之间的同质性统计项，标注为代码“N”，而执行了同一项目的机构之间的同质性统计项则标注为代码“Y”。由于代码 N 在代码 Y 之前，因此，利用 nodematch 命令进行模型估计，两个均未执行某一项目的机构

表 3.8 基于地域与执行项目差异化同质性的同质性模型

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodecov("popmil") + nodefactor("years") + nodematch("hivscreen",
    diff = T) + nodematch("nutrition", diff = T) + nodematch("state")

Iterations:  20

Monte Carlo MLE Results:
                         Estimate Std. Error MCMC %      Lower         OR   Upper  p-value
edges                  -9.548e+00 1.131e-01    NA  5.714e-05  7.132e-05   0.000 < 1e-04 ***
nodecov.popmil          3.306e-01 2.009e-02    NA  1.338e+00  1.392e+00   1.448 < 1e-04 ***
nodefactor.years.1      1.757e-01 4.748e-02    NA  1.086e+00  1.192e+00   1.308 0.000215 ***
nodefactor.years.2      3.234e-01 4.445e-02    NA  1.267e+00  1.382e+00   1.508 < 1e-04 ***
nodefactor.years.3      3.468e-01 4.236e-02    NA  1.302e+00  1.415e+00   1.537 < 1e-04 ***
nodematch.hivscreen.N  -2.571e-02 6.203e-02    NA  8.630e-01  9.746e-01   1.101 0.678461
nodematch.hivscreen.Y   4.490e-01 4.990e-02    NA  1.421e+00  1.567e+00   1.728 <1e-04  ***
nodematch.nutrition.N   1.013e-02 8.302e-02    NA  8.585e-01  1.010e+00   1.189 0.902902
nodematch.nutrition.Y   2.495e-01 4.861e-02    NA  1.167e+00  1.283e+00   1.412 < 1e-04 ***
nodematch.state         6.313e+00 8.440e-02    NA  4.675e+02  5.516e+02 650.802 < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   19457  on 822393  degrees of freedom
          Deviance: 1120635  on     10  degrees of freedom

AIC: 19477    BIC: 19593
```

表 3.9　仅保留一种分类情形的差异化同质性统计项的同质性模型

```
==========================
Summary of model fit
==========================

Formula:    lhds ~ edges + nodecov("popmil") + nodefactor("years") + nodematch("hivscreen",
    diff = T, keep = 2) + nodematch("nutrition", diff = T, keep = 2) +
    nodematch("state")

Iterations:  20

Monte Carlo MLE Results:
                       Estimate Std. Error MCMC %      Lower        OR   Upper  p-value
edges                 -9.556e+00 1.100e-01    NA  5.707e-05 7.080e-05  0.000  < 1e-04 ***
nodecov.popmil         3.310e-01 2.005e-02    NA  1.339e+00 1.392e+00  1.448  < 1e-04 ***
nodefactor.years.1     1.756e-01 4.748e-02    NA  1.086e+00 1.192e+00  1.308  0.000216 ***
nodefactor.years.2     3.238e-01 4.443e-02    NA  1.267e+00 1.382e+00  1.508  < 1e-04 ***
nodefactor.years.3     3.463e-01 4.233e-02    NA  1.301e+00 1.414e+00  1.536  < 1e-04 ***
nodematch.hivscreen.Y  4.587e-01 4.352e-02    NA  1.453e+00 1.582e+00  1.723  < 1e-04 ***
nodematch.nutrition.Y  2.496e-01 4.504e-02    NA  1.175e+00 1.284e+00  1.402  < 1e-04 ***
nodematch.state        6.310e+00 8.411e-02    NA  4.666e+02 5.502e+02 648.811 < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   19457  on 822395  degrees of freedom
          Deviance: 1120635  on      8  degrees of freedom

AIC: 19473    BIC: 19566
```

(N-N)之间的同质性项位于第一位，而两个执行了同一项目的机构间的同质性项(Y-Y)将置于后面。如果仅希望保留两个执行了同一项目的机构间的同质性项，则可以通过在"Command 25"(命令25)中增加 keep=2 参数来实现。

计算同质性与差异化同质性项的变化统计与之前所提到的主效应项的计算方法十分类似，稍有不同的地方是，同质性项的变化统计关注的是二元组，因此仅会产生两个可能的变化统计值。根据亨特、古德鲁及其同事(2008)和古德鲁及其同事(2009)所提供的方法，变化统计可以表示为：

同质性的变化统计：

$$\delta_{\text{hom}}=\begin{cases}1\text{，如果 } i \text{ 与 } j \text{ 对分类协变量具有相同的值}\\0\text{，其他}\end{cases}$$

差异化同质性的变化统计：

$$\delta_{\text{diff}}=\begin{cases}1\text{，如果 } i \text{ 与 } j \text{ 对分类协变量的某一种分类具有相同的值}\\0\text{，其他}\end{cases}$$

通过计算图3.7中两个地方卫生机构建立沟通关系的概率，就可以解释如何使用二元组层次的统计项。每一个地方卫生机构都含有诸如领导履职经验、辖区人口数量等个体特征。除此之外，在二元组层面上，两个地方卫生机构之间在地域(例如，两个机构都位于密苏里州)和执行营养项目方面也存在一致性特征，但在艾滋病筛查项目上则没有体现出一致性特征(也许较大的机构提供了艾滋病筛查项目，而小型机构不提供该项目)。该模型有10个统计项；它会先显示完整的模型，但只有那些适用于相关地方卫生机构的统计项会显示替代值。最终，模型计算了两个地方卫生机构建立沟通

关系的概率(出于简化的便利,同质性被缩写为"Hom"):

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}\begin{pmatrix} \theta_{\text{edges}}\delta_{\text{edges}} + \theta_{\text{popmil}}\delta_{\text{popmil}} + \\ \theta_{3-5\text{years}}\delta_{3-5\text{years}} + \theta_{6-10\text{years}}\delta_{6-10\text{years}} + \\ \theta_{>10\text{years}}\delta_{>10\text{years}} + \theta_{\text{HIVHom}}\delta_{\text{HIVHom}} + \\ \theta_{\text{NutritHom}}\delta_{\text{NutritHom}} + \\ \theta_{\text{StateHom}}\delta_{\text{StateHom}} \end{pmatrix}$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}\begin{pmatrix} -9.56 * \delta_{\text{edges}} - 0.33 * \delta_{\text{popmil}} + 0.32 * \delta_{6-10\text{years}} \\ +0.25 * \delta_{\text{NutritHom}} + 6.31 * \delta_{\text{StateHom}} \end{pmatrix}$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(-9.56*1 - 0.33*2.1 + 0.32*1 + 0.25*1 + 6.31*1)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = 0.033$$

同质性模型较主效应模型在沟通关系建立的概率上有较大幅度的提升,主要原因是处于同一个州这一因素导致系数变大。根据这一模型,这两个地方卫生机构之间建立沟通关系的概率提升至了 3.3%(图 3.8)。

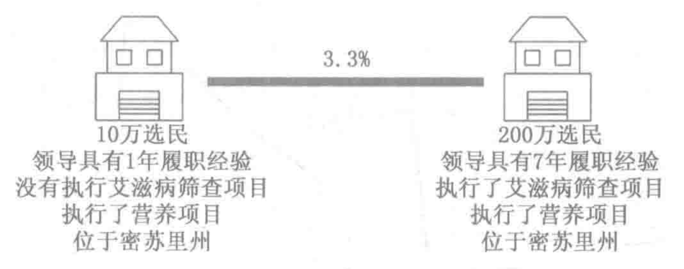


图 3.8 基于差异化同质性模型对地方卫生机构建立沟通关系概率的估计

虽然模型中有许多预测变量已经表现出了统计上的显著性，并且与前期根据探索性分析所观察出的模式一致，对于模型的效度已经进行了初步的检验，但更为重要的任务是，我们还需要更为系统地检验：究竟估计模型能够在多大程度上反映观察网络的结构特征。

模型拟合

统计网络模型中有几种检验模型拟合优度的方法。其中，最简单的方法是：将模型的对数似然估计结果(log-likelihood, LL)与对应的离差的测量结果(deviance, −2LL)、赤池信息准则(Akaike information criterion, AIC)以及贝叶斯信息标准(Bayesian information criterion, BIC)进行比较。最大似然估计值通过观测网络中 Y_{ij} 真实发生联系的概率与 Y_{ij} 的期望概率之间的差异值求和获得的(Field, 2009)。

$$\text{log-likelihood} = \sum_{i=1}^{N} [Y_{ij} \ln(P(Y_{ij})) + (1 - Y_{ij}) \ln(1 - P(Y_{ij}))] \quad [3.3]$$

简而言之，LL 是对网络中二元组形成连接关系的期望概率和实际发生概率之间差值乘积的求和。

例如，考虑如下一种情形，当两个地方卫生机构之间建立了一个沟通关系($Y_{ij}=1$)，如图 3.6。之前，在差异化同质性模型中估计两个地方卫生机构之间存在沟通关系的概率是 0.23%($P(Y_{ij})=0.0023$)。因此，如果在地方卫生机构网络中，两个地方卫生机构之间存在沟通关系，那么，该二元组对于模型的对数似然估计值的贡献是：

$$Y_{ij}\ln(P(Y_{ij}))+(1-Y_{ij})\ln(1-P(Y_{ij}))$$
$$1*\ln(0.0023)+(1-1)*\ln(1-0.0023)=-6.07$$

而如果两个地方卫生机构之间没有建立沟通关系($Y_{ij}=0$),那么,就说明该二元组对模型对数似然估计值的贡献为:

$$Y_{ij}\ln(P(Y_{ij}))+(1-Y_{ij})\ln(1-P(Y_{ij}))$$
$$0*\ln(0.0005)+(1-0)*\ln(1-0.0005)=-0.0023$$

可见,预测的地方卫生机构之间建立沟通关系的概率十分低(0.23%),因此,两个地方卫生机构之间建立了沟通关系对于对数似然值的贡献度就要远大于没有建立沟通关系的二元组的贡献度。因此,当预测概率与观测网络中的实际概率相差较大时,对数似然估计的值就会增大;预测的概率与实际概率相差越多,对数似然估计值就越高。从概念上而言,在对模型失拟进行量化方面,对数似然估计十分类似于线性回归模型中的残差平方和(Field, 2009)。由于对数似然估计值(LL)常常是负值,直观地进行比较十分困难,因此,为了克服这种困难,研究人员采用离差(−2LL)方法来取代对数似然估计方法(LL)。离差方法仅仅是在对数似然值的基础上乘以−2,但却能够确保结果是正值。离差方法也被认为是一种对模型失拟进行检验的方法,离差越大,模型失拟的程度就越高。

离差方法可以用于对具有嵌套关系的两个网络规模不同的模型直接进行比较,从而判断在模型的拟合优度方面,规模较大的模型是否要显著地优于规模较小的模型。两个嵌套模型所对应的离差值之间的差异服从一个卡方分布(chi-squared distribution),其自由度等于两个模型在参数数量上的差异。这种情况下,如果主效应模型的离差为1 103 927且自由度为

5，而差异化同质性模型的离差为 1 120 635 且自由度为 10，那么两者的差异是 16 708，有 10－5＝5 个自由度。我们将这个值与卡方分布进行比较，发现 p 值小于 0.0001，说明在模型拟合方面，差异化同质性模型显著地优于主效应模型（$\chi^2(5)=16\,708$；$p<0.0001$）。因此，为模型增加同质性统计项的做法显著地改进了模型的拟合度。

与离差（－2LL）方法不同，赤池信息准则（AIC）和贝叶斯信息标准（BIC）是另外两种评价模型拟合效果的方法。通常情况下，模型所包含的参数越多，离差的值就会越小；AIC 和 BIC 方法恰恰是考虑到这一点，通过惩罚那些包含了过多参数但并没有解释足够丰富信息模型的做法*，因此，AIC 和 BIC 方法被认为是进行模型拟合效果评价的更好方法（Akaike，1973；Schwarz，1978）。根据这种思路，这两种评价方法给出了一种类似线性回归中校正判定系数（adjusted R^2）的方法，然而，这些方法本身是无法直接解释的，但可以用来进行模型间的比较。在公式 3.4 中，p 代表模型中参数的数量，N 表示样本规模。

$$\begin{aligned}\text{AIC}&=\text{Deviance}+2p\\\text{BIC}&=\text{Deviance}+p*\ln(N)\end{aligned}\qquad[3.4]$$

AIC 和 BIC 方法较离差方法更为灵活，因为它们可以用于比较非嵌套模型。在地方卫生机构的各种模型中，零模型的 AIC 是 36 367，主效应模型的 AIC 是 36 176，差异化同质性模型的 AIC 则下降到 19 477，而改进的差异化同质性模型的 AIC 又下降了一些，到达 19 473。因此，根据 AIC 的结果，

---

* 过拟合现象。——译者注

改进的差异化同质性模型是目前拟合优度最佳的模型。

上面提到这些针对拟合优度评价的方法是适合于以独立性假设为基础的观测网络数据的。因此，要想评价一个指数随机图模型在多大程度上能够表征观测网络的结构特征，另外一些方法通常被认为是更适合进行网络结构特征的评价。至此，由于零模型、主效应模型、同质性模型是符合二元组的独立性假设的，因此，离差方法、AIC 和 BIC 方法对于模型评价还是有效的；但当模型更加复杂时，如包括了二元依赖性乃至其他更高阶的依赖性等假定，那么，我们就需要采用基于仿真的模型拟合优度评价方法。

一种简单的采用模型仿真对拟合优度进行评价的方法是：首先，基于模型对单个网络进行仿真，并比较该仿真网络与观测网络的特征差异，接下来，利用现已构建的每一种模型进行网络仿真，通过这种方法“Command 26”（命令 26）检验并比较每一种模型的构建效果。

可见通过仿真获得的网络与观测网络之间存在一定的差异（参见表 3.10），例如，与经仿真所获得的若干网络（参见表 3.10 第二行至第六行的五种网络）相比，观测的地方卫生机构网络（参见表 3.10 中第一行高亮部分）在网络中具有孤立节点数量（degree＝0）和三角形数量（triangle）上表现出了明显优势。虽然，这些仿真网络在反映网络结构的特征方面还有很大的改进空间，但需要重点关注的是：仿真网络通过不断增加统计项做法将核心的社会过程纳入到模型的考量范围中来，从而使仿真网络不断地逼近观测网络。以三角形数量为例，前面的五种模型均没有包含三角形数量统计项，但是，仿真网络的三角形数量从最初简单随机图模型（零模型）

中的 17 个迅速增加到差异化同质性模型中的 1 249 个。*

表 3.10　地方卫生机构网络与仿真网络在边数量、节点中心度(0—5)以及三角形数量上的比较

	edges	degree0	degree1	degree2	degree3	degree4	degree5	triangle
lhds	2708	58	117	182	223	226	172	1437
Null	2647	18	97	159	243	276	196	17
Main effects	2660	29	95	166	243	246	202	32
Homophily	2704	48	127	149	234	244	168	1223
Diff homophily	2707	45	125	169	224	231	174	1249
Diff homophily 2	2713	48	112	182	222	233	170	1249

增加如表 3.10 所示的仿真可以为我们的模型拟合效果评价提供额外参考。如果我们以一个模型为基础进行 10 次(也可以是其他数量)网络仿真,那么我们就可以利用"Command 27"(命令 27),比较这一组仿真网络与观测网络的统计结果差异。例如,命令 27 所导出的部分结果显示,nodefactor.popmil(辖区人口)网络统计值的范围是从 1 285.733 至 1 351.685,而该数据正是以之前改进的差异化同质性模型为基础进行网络仿真获得的 10 个网络。命令 27 导出的部分结果如下:

```
Stored network statistics:
      edges nodecov.popmil
 [1,]  2701       1333.655
 [2,]  2689       1315.555
 [3,]  2704       1338.370
 [4,]  2710       1351.434
 [5,]  2711       1345.763
 [6,]  2722       1347.097
 [7,]  2710       1351.685
 [8,]  2720       1329.238
 [9,]  2717       1287.713
[10,]  2719       1285.733
```

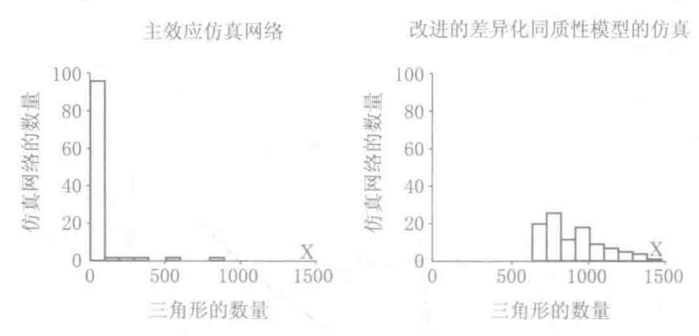
* 通过上面例子,可以了解模型仿真是如何通过不断纳入网络的结构特征从而实现逼近真实观测网络的目的的。——译者注

一个连续性主效应(nodecov)统计项等于网络中连线两端的节点所对应的变量值的总和。我们可以通过“Command 28”(命令28)获得地方卫生机构网络中辖区人口的观测值:

```
nodecov.popmil
        1345.815
```

在所观测的地方卫生机构网络中,辖区人口的网络统计结果是1 345.815。而以改进的差异化同质性模型为基础的仿真网络,其网络所显示的辖区人口的统计范围是1 285.733—1 351.685。

我们可以以主效应模型和改进的差异化同质性模型为基础分别拟合100个(或者是其他任何数量的)网络,然后利用“Commands 29—31”(命令29—31),观察这组仿真网络集合在三角形数量分布上与观测网络的差异,结果参见图3.9。通过上述比较,我们可以理解主效应模型和差异化同质性模型如何反映地方卫生机构网络的传递性特征(Goodreau et al., 2008)。



依据主效应模型以及改进的差异化同质性模型所构建的100个仿真网络的三角形分布。其中,X标记了地方卫生机构网络观测到的三角形数量。

图3.9

由主效应模型所产生的100个仿真网络中，仅有5个仿真网络包含了超过100个三角形；而采用改进的差异化同质性模型进行仿真的网络则每一个网络都产生了500个以上的三角形。由此可见，差异化同质性模型是对主效应模型的重大改进。尽管如此，改进的差异化同质性模型对所观测的地方卫生机构网络中的三角形数量还是低估了。所观测的地方卫生机构网络一共包含1 437个三角形（在图3.9中标注为X）。这样的结果说明：上述任意一个模型均没有能够很好地表征地方卫生机构网络的传递性特征。

将模型仿真纳入到网络拟合优度的评价过程，能够更好地比较仿真网络与观测网络在网络特征上表现出来的差异。目前，R中的ergm包已经将两种网络（仿真网络和观测网络）的度分布、边共享伙伴和二元组共享伙伴测量，嵌入到网络拟合优度的评价过程中来。利用这个步骤所获得的结果，拟合优度的评价可以采取如下两种方式。首先，比较仿真网络与观测网络在每一个网络统计项上频次的差异"Command 32"（命令32）。

表3.11　差异化同质性模型的拟合优度评价结果

Goodness-of-fit for degree

	obs	min	mean	max	MC p-value
0	58	32	58.78	78	0.96
1	117	111	134.76	176	0.12
2	182	159	190.23	217	0.62
3	223	177	205.45	233	0.16
4	226	151	188.28	237	0.08
5	172	120	152.33	196	0.28
6	104	86	116.71	144	0.22
7	67	59	84.98	109	0.04
8	35	29	55.45	74	0.02

续表

```
9    25   20   37.61   60         0.08
10   26   15   24.70   42         0.80
11   14   8    14.97   26         0.84
12    8    2    9.03   17         0.88
13    6    1    4.43   10         0.54
14    8    0    1.96    7         0.00
15    4    0    1.05    4         0.08
16    3    0    0.50    3         0.02
17    1    0    0.50    3         0.76
18    1    0    0.28    3         0.50
19    1    0    0.23    2         0.44
20    1    0    0.22    1         0.44
21    0    0    0.11    1         1.00
22    1    0    0.11    2         0.20
23    0    0    0.10    1         1.00
24    0    0    0.11    1         1.00
25    0    0    0.07    1         1.00
26    0    0    0.04    1         1.00
29    0    0    0.01    1         1.00

Goodness-of-fit for edgewise shared partner
       obs  min     mean  max MC p-value
esp0   696  923  1652.45 1808       0.00
esp1   750  647   723.50  805       0.56
esp2   630  153   232.33  578       0.00
esp3   382   33    63.65  322       0.00
esp4   156    5    15.72  109       0.00
esp5    56    5     4.50   47       0.00
esp6    25    0     1.04   18       0.00
esp7     8    0     0.32    7       0.00
esp8     3    0     0.09    2       0.00
esp9     0    0     0.05    1       1.00
esp10    1    0     0.03    1       0.06
esp11    1    0     0.03    1       0.06

Goodness-of-fit for dyadwise shared partner
         obs     min      mean     max MC p-value
dsp0  813034  811054  811708.89  812789       0.00
dsp1    6329    6795    8477.22    9143       0.00
dsp2    1928    1543    1767.79    1929       0.02
```

续表

dsp3	732	270	367.36	649	0.00
dsp4	253	40	66.60	204	0.00
dsp5	80	3	12.55	71	0.00
dsp6	33	0	2.03	27	0.00
dsp7	9	0	0.36	7	0.00
dsp8	3	0	0.09	2	0.00
dsp9	0	0	0.05	1	1.00
dsp10	1	0	0.03	1	0.06
dsp11	1	0	0.03	1	0.06

这些统计摘要表均包含5列信息：obs、min、mean、max以及MC p value(表3.11)。统计表的第一列列出了每个具体的统计项(如degree、ESP、DSP)；obs列显示的是地方卫生机构网络中各统计项所对应的节点的数量；min显示的是当度、边共享伙伴或者二元组共享伙伴数量确定时，基于不同模型所建立的仿真网络中的最小节点数量；mean显示的是当度、边共享伙伴或者二元组共享伙伴数量确定时，基于不同模型所建立的仿真网络中的节点的平均数；max显示的是当度、边共享伙伴或者二元组共享伙伴数量确定时，基于不同模型所建立的仿真网络中的最大节点数量；MC p value列显示的是仿真网络统计值与观测网络统计值至少同样极端的比率。如果MC p 值较大，说明仿真网络与观测网络相应的网络特征十分相似(或者说，不存在显著差异)，而如果MC p 值较小则说明观测网络与拟合网络在统计频次上的差异；于是，MC p 值小于0.05被解释为观测网络与仿真网络之间存在显著差异；这也表明仿真模型没有很好地拟合真实的观测数据。表3.11阴影部分所有的 p 值都小于0.05，表明仿真网络并没有能够很好表征观测网络的结构特征。

表3.11中的第一个表是针对网络中心度特征的拟合优

度评价，我们可以根据该表第一行的值发现：真实的地方卫生机构网络中包含 58 个孤立节点（即度为 0 的节点）；相比而言，仿真网络中包含孤立节点的平均数为 58.78，而孤立节点的数量分布范围是 32 至 78。从仿真网络的节点平均数和分布范围，我们发现在获取观测网络的这一特征方面，仿真网络的效果很好。接下来，当度显示为 0 时，拟合优度评价所对应的 MC p 值是 0.96，说明当网络的度固定为 0 时，观测网络和仿真网络之间在节点数量方面并没有显著差异。仿真网络能够在大多数度值情况下，很好地获取网络节点的数量。在观测网络中，度值为 1 的节点共有 117 个；在对应的仿真网络集合中，度值为 1 的平均节点数为 134.76 个。MC p 值为 0.12，显示该仿真网络很好地表征了观测网络，说明所观测网络与网络之间并不存在显著的差异。通常而言，在这些表格中，有越多小于 0.05 的 p 值，说明网络拟合越好。

同时，表 3.11 显示了边共享伙伴与二元组共享伙伴存在部分失拟的问题。基于改进的差异化同质性模型产生的仿真网络仅包含了较少的边共享伙伴数量和少数的二元组共享伙伴数量（DSP＝9，DSP＝10，DSP＝11）*。鉴于边共享伙伴和二元组共享伙伴指标是测量网络传递性的指标，那么，仿真网络在上述两个指标测量上所表现出的拟合效果不好的现象，与图 3.9 中所反映出的三角形数量缺失的现象是一致的。这进一步说明了改进的差异化同质性模型并没有很好地获取观测网络的传递性特征。

值得注意的是：表中并没有显示每一个网络统计项所有

---

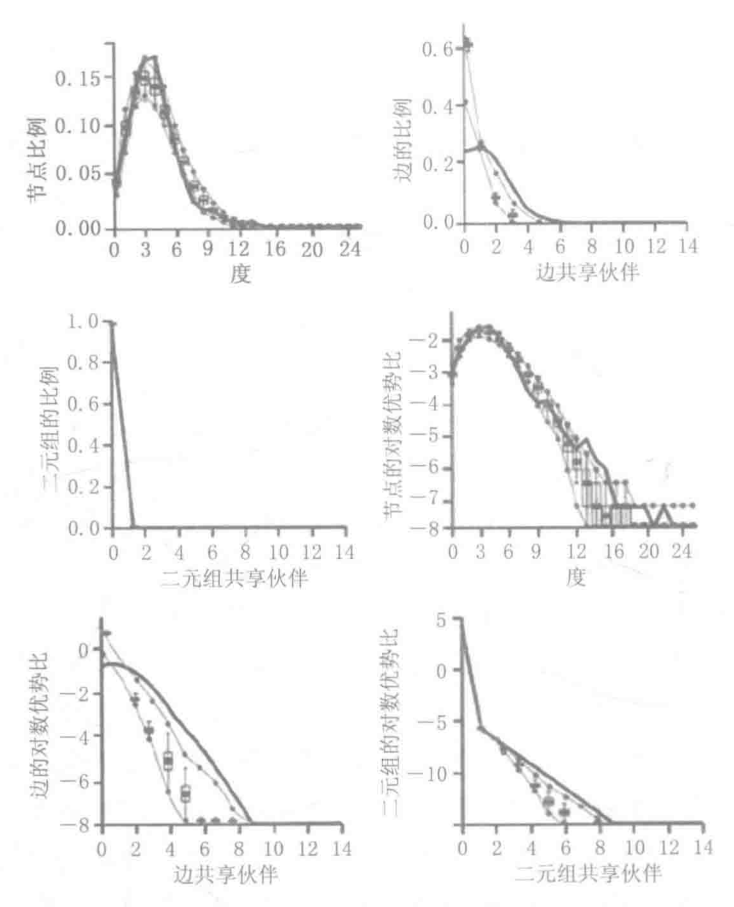
* 不能很好地模拟观测网络的边共享伙伴和二元组共享伙伴的结构特征。——译者注

可能的值。例如，在地方卫生机构网络中，网络具有1 283个成员，每个节点都可以与其他节点建立最多1 282项沟通关系，但结果显示中心度的范围仅为0到29；如果有必要，可以采用"Command 33"（命令33）查看每一个表中所有行的值，但那样导出的结果会特别长。

R中statnet包中的拟合优度评价过程还包含了三元组（triangle census）和捷径距离（geodesic distance）测量等选项。这些测量选项并没有包含在statnet包内嵌的拟合优度评价过程中，如果想采用上述测量选项去对网络仿真效果进行评价，则可以采用单独的仿真步骤。例如，可参见"Commands 29—31"（命令29—31）对于三角形数量的测量过程。

除了对每一个网络统计项的观测值和拟合值进行频次上的比较之外，拟合优度评价的过程也可以采用可视化图形观察的方法。图形观察方法不再是对每一个网络统计项的频次进行比较，而是对比仿真网络和观测网络在具有相同特征节点数量比例上的差异，参见"Command 34"（命令34）。另外，当绘图的参数设置被更改时，图形观察方法可以用来比较观测网络中每一个参数的对数优势比以及仿真网络中对数优势比的范围。例如，之前的表已经比较了观测网络和仿真网络在孤立节点数量上的差异，其中，观测网络的孤立节点数为$n=58$，而仿真网络中孤立节点的取值范围是$n=$ 32—78；因此，就可以通过绘制图形的方法比较观测网络中孤立节点的比例（4.5%）与仿真网络中孤立节点比例的差异（参见图3.10的上面一行）。图3.10的下面一行比较了观测网络中各统计项的对数优势比与仿真网络中所对应统计项的对数优势比范围之间的差异。由于通过对数转化形式的

视图模式更易于判别模型拟合的程度，因此，对数优势比方法也被沿袭到了拟合优度评价的绘图中来。



针对差异化同质性模型的仿真网络进行拟合优度评价。其中，黑色线代表观测值；灰色线以及箱型图代表仿真网络的测量结果；前三张图显示的是比例，后三张图显示的是对数优势比。

图 3.10

在图 3.10 中，粗黑色线代表了地方卫生机构观测网络的

测量结果，灰色线则代表了仿真网络在95%的置信区间时的测量结果。当黑色线落在灰色线条之间时，就说明仿真网络能够较好地代表观测网络的结构特征。在本例中，根据模型仿真，我们发现仿真网络能够较好地解释度中心度、二元组共享伙伴等特征，但是，边共享伙伴统计项则在拟合评价过程中存在一些问题。

需要注意的是，原本在模型拟合优度的评价方面，表3.11和图3.10应该表现为完全一致的，然而，该仿真网络的两种表现形式却显示了极大的差异。表3.11比较了观测网络和仿真网络在特定统计值上的节点频次情况，而图3.10则是比较了具有特定统计值的观测网络和仿真网络所对应的节点比例（或对数优势比）情况。

截至目前，我们已经描述了三种进行拟合优度检验的策略：

1. 利用模型统计摘要表中所包含的AIC和BIC，可以对模型的拟合优度进行评价，AIC和BIC的值越低，说明仿真网络的拟合优度越好。

2. 以一个或多个模型为基础进行仿真，通过比较仿真网络与观测网络特征上的差异，理解所建构的模型究竟在多大程度上能够表现观测网络的基本特征（例如，中心度和三角形数量特征等）。

3. R的statnet包中拟合优度评价程序为比较观测网络和仿真网络的结构特征的测量提供了一系列的统计表和可视化工具，如中心度、距离、边共享伙伴、二元组共享伙伴、三元组等。统计摘要表提供p值作为判别手段，统计图则利用置信区间作为判断手段，这些方法都能够对观测网络与仿真

网络的测量结果是否符合同一分布进行评价。这些内嵌在statnet包中的拟合优度评价与第二种策略相类似。

虽然拟合优度评价的各种评测结果显示，改进的差异化同质性模型在拟合优度的诸多方面较主效应模型已经有了较大程度的提升，但改进的差异化同质性模型并没有能够很好地拟合真实观测网络。这对于二元独立性模型是较为常见的现象：虽然最大似然估计过程找到了最有可能复制观测网络的模型，但这个可能性仍然是比较低的（Hunter & Goodreau et al., 2008）。拟合优度测量的结果显示：主效应以及同质性项不能准确地把握传递性的特征。因此，如果能考虑增加一些涉及网络内在分布特征以及复杂依赖关系的统计项，或许能帮助我们改进模型拟合。

增加依赖关系项

为了解释观测网络存在复杂的依赖关系，斯尼德斯和他的同事（2006）提出三个统计项，后来亨特和汉考特（2006）为了简化说明，对此作了修订。这三个统计项是：几何加权度（GWD）、几何加权边共享伙伴（GWESP）和几何加权二元组共享伙伴（GWDSP）。上述三个统计项从观测网络内依赖关系间的复杂模式出发，考察网络的度分布及传递性特征（关于这些统计项的更多信息可参见第2章）。目前，这些经过修订的统计项作为依赖性模型评价的工具已经被纳入到了statnet包中。

之前在二元独立性模型上所采用的最大似然估计方法，会由于计算量过于庞大而难以在二元依赖性模型复制。如，

计算公式 2.6 的常数项往往需要对网络中所有可能的网络配置结果进行汇总，而网络数量有 $2^{\binom{n}{2}}$ 个，于是，对于一个仅拥有 9 个节点的网络而言，其网络的配置就会产生 68 719 476 736 种情形(Cranmer & Desmarais，2011)。因此，二元依赖性模型需要使用马尔科夫链蒙特卡洛(Markov chain Monte Carlo，MCMC)参数估计算法来计算一个近似的对数似然(log-likelihood)结果(Snijders，2002)。* 默认条件下，最大伪似然估计(maximum pseudolikelihood)方法被用于判别模型估计的初始值；接下来，MCMC 算法从所有可能实现的网络中选择一个网络，从该网络中随机地选择一个二元组或者多个二元组，对一个二元组或者多个二元组实施从 0 至 1 或者从 1 到 0 的转换，通过比较切换后的网络与切换前的网络，观察哪一个网络会有更好的拟合效果；接下来，算法需要考虑是接受转换后的新网络，还是保留转换前的网络并继续下一轮随机二元组选择和转换。这种“提出(propose)——比较(compare)——决定(decide)”的过程会被重复多次，直至整个 MCMC 链全部进行完毕(Morris et al.，2008)。

第 2 章曾经讨论过，即使模型已经纳入了同质性或者其他统计项，模型仍可能会存在近似退化问题，该问题说明所构建模型尚未完全获得观测网络的结构特征。网络建模中近似退化现象常常表现为：基于模型仿真的网络或者近乎空图(网络中各节点完全不联系)或者近似为全图(网络中各节点全部相连)(参见 Robins et al.，2007 中的图 1，就是一个比较典型的图形实例)。增加几何统计项的初衷正是要解决在

* 而不采用原先的最大似然估计法。——译者注

早期网络建模过程中遇到的近似退化问题，因此，这些几何参数（GWD、GWDSP、GWESP）也就成为了我们模型构建中的一部分内容。

我们除了考虑模型需要纳入什么之外，还应该考虑如何估计模型。具体而言，我们需要考虑采用一些额外的步骤来降低模型无法收敛的概率，包括选择充分的 MCMC 样本规模、老化次数（burn-in）以及间隔（interval）等（Goodreau et al., 2008; Morris et al., 2008）。样本规模控制了整个 MCMC 链中网络分析样本的数量（之前的章节中描述过马尔科夫链的长度）；老化次数则是指我们在选择网络最初的网络样本规模时，需要先排除多少个网络*；间隔则用来确定两个样本之间所跨越的样本数量。如果一个网络模型显示出了近似退化的迹象，那么，我们可以通过增大上述参数的设定并重新对模型进行评估来帮助模型获得收敛。需要注意的是，大多数用户都会发现，增加 MCMC 的样本规模会导致 R 的模型估计时间呈小时级的增长。上述三个参数设置都可以在 control.ergm 命令中设置。为了让上述模型的评价结果可重复，我们可以在命令中加入一个种子值（seed value），这样模型每次都能被指定从同一地点开始。这个种子值的设置命令也被增加到 control.ergm 中来了。

根据古德鲁及其同事的建议（2008；参见第 2 章），这里 α 可以首先选择 0.1，然后逐步增加直至对数似然值不再增长为止。因此，对包含几何统计项的模型进行拟合优度评价时，α 值的设定也是从 0.1 开始（命令 35）。需要读者注意的

---

* 以确保网络变化过程趋于稳定。——译者注

是，我们所使用的命令，每一个模型的计算都需要大量的时间才能使模型收敛（小时级别）。在R中，虽然在某些条件下并行处理是允许的，但由于R默认地限定无论用户电脑中包含几核处理器仅允许使用单核，于是，对于大多数用户而言，往往仅能使用单核。没有先进的计算能力，模型的计算速度是难以提升的。针对上述问题，statnet 包的开发团队十分努力地增加并行处理功能，参见 ergm 包中 R 的文档中关于 ergm-parallel 的记录（http://cran.r-project.org/web/packages/ergm/ergm.pdf）。

对拟合优度评价采用可视化展现的方式对于依赖性模型检验而言可能更为直观，同时值得注意的是，AIC 和 BIC 的拟合优度测量结果也经常与可视化测量结果相互印证，并且在 α 值的选择过程中起到迅速比较的作用。利用“Command 35”（命令 35）可以分别产生模型在 α 值为 0.1、0.2、0.3、0.4、0.5、0.6、0.7、**1** 以及 1.1 条件下的 AIC 值，分别为 18 019、17 943、17 875、17 814、17 759、17 732、17 700、**17 660**、17 667。对 BIC 的检验也显示出一条类似的轨迹。根据上述结果可知，当 $\alpha=1$ 时，模型拟合优度的评价值 AIC 和 BIC 是最好的（参见上面 AIC 数值中加粗的数字）。于是，我们以改进的差异化同质性模型的统计项为基础，增加三个当 $\alpha=1$ 时的几何加权统计项，用来估计一个新的二元依赖性模型（参见表 3.12）。

在依赖性模型所包含的统计项中，有些系数是正向且显著的，说明地方卫生机构建立沟通关系的概率与地域同质性、执行项目的同质性、辖区人口数量、领导履职年限、几何加权度和几何加权边共享伙伴有正向关系。模型中，几何统

表 3.12　以依赖性假设为条件模型(指定 $\alpha=1$)

```
==========================
Summary of model fit
==========================

Formula:   lhds ~ edges + nodefactor("years") + nodecov("popmil") + nodematch("hivscreen",
    diff = T, keep = 2) + nodematch("nutrition", diff = T, keep = 2) +
    nodematch("state") + gwdegree(1, T) + gwesp(1, T) + gwdsp(1,
    T)

Iterations:  20

Monte Carlo MLE Results:
                          Estimate Std. Error   MCMC %      Lower         OR     Upper p-value
edges                   -1.007e+01 3.621e-01  7.900e+01  2.090e-05  4.250e-05   0.000 < 1e-04 ***
nodefactor.years.1       1.426e-01 4.983e-02  1.600e+01  1.046e+00  1.153e+00   1.272 0.00420 **
nodefactor.years.2       2.541e-01 4.691e-02  1.800e+01  1.176e+00  1.289e+00   1.413 < 1e-04 ***
nodefactor.years.3       2.969e-01 4.795e-02  2.100e+01  1.225e+00  1.346e+00   1.478 < 1e-04 ***
nodecov.popmil           2.005e-01 2.304e-02  1.000e+00  1.168e+00  1.222e+00   1.278 < 1e-04 ***
nodematch.hivscreen.Y    1.750e-01 4.200e-02  1.000e+00  1.097e+00  1.191e+00   1.293 < 1e-04 ***
nodematch.nutrition.Y    1.897e-01 4.208e-02  1.300e+01  1.113e+00  1.209e+00   1.313 < 1e-04 ***
nodematch.state          5.018e+00 1.210e-01  4.700e+01  1.192e+02  1.511e+02 191.504 < 1e-04 ***
gwdegree                 1.851e+00 2.193e-01  6.800e+01  4.144e+00  6.369e+00   9.790 < 1e-04 ***
gwesp.fixed.1            9.588e-01 2.941e-02  2.900e+01  2.463e+00  2.609e+00   2.763 < 1e-04 ***
gwdsp.fixed.1           -3.936e-02 1.377e-02  6.400e+01  9.358e-01  9.614e-01   0.988 0.00427 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual  Deviance:   17638  on 822392  degrees of freedom
           Deviance: 1122455  on     11  degrees of freedom

AIC: 17660    BIC: 17788
```

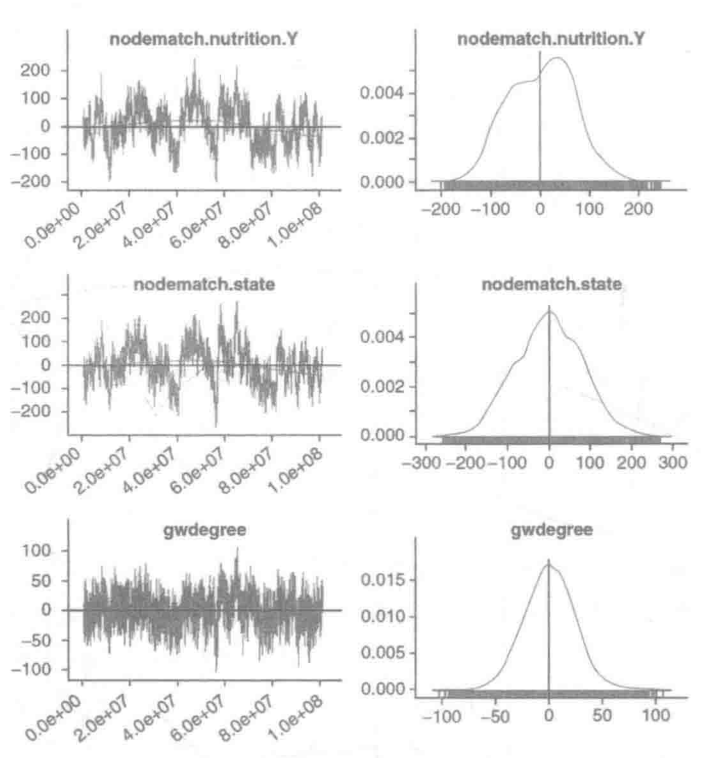
计项的系数显示为正向且显著表明：假定网络其他条件保持不变，同时给定网络中的度分布、边共享伙伴和二元组共享伙伴，那么两个地方卫生机构之间建立沟通关系的概率要高于其随机建立沟通关系的概率。下一部分我们将通过展现计算过程解释这些几何统计项是如何对关系形成的概率产生影响的。

MCMC 模型诊断

除了利用上面讨论的这些策略检测模型的拟合优度之外，模型诊断（model diagnostics）也能够辅助判断估计算法是否已经收敛还是存在近似退化问题，进而判断究竟是模型本身还是模型估计设置条件需要进行调整。模型诊断的第一项策略是检验在程序迭代过程中对数似然估计值的变化情况，我们可以通过在模型命令中输入 verbose = T，选择将该模型估计过程显示出来，参见“Command 35”（命令 35）。对数似然值的增量显示了模型初始值与经过多次迭代后拟合值之间的差距，较大的改进数值说明模型初始值的作用完全消失了。因此，当拟合过程中任意一次迭代产生的对数似然值（LL）变化值超出原始值的 20 倍时，MCMC 的运算就会停止，因为上述结论显示可能存在一个近似退化的模型，或者说模型初始值与模型最终估计值之间存在过大的差异。通常而言，迭代过程中的变化应该是比较小的，而且伴随着迭代次数的增加，这种变化会逐渐减小。

除了检验对数似然值（LL）的变化值之外，对 MCMC 诊断的图形化展示也会是非常有效的（Command 36）。MCMC

诊断图显示了在模型最后迭代的阶段，模型呈现的状态（参见图 3.11）；图 3.11 左侧的绘图，以模型中的每一个统计项为单位，利用 MCMC 链作一个时间序列来展示统计项的变化情况，右边的绘图则显示了对应 MCMC 链的直方图（Goodreau et al.，2008）。



当 $\alpha=1$ 时，针对依赖性模型中的多个统计项进行 MCMC 诊断。

图 3.11

如果模型能够收敛，那么，模型中每一个统计项的图将

会表现为以0为中心随机变化，这里0代表观测网络对应统计项的统计值。本例中，大多数统计项的图表都是围绕0随机变化的，除了执行营养项目同质性项和其他少数统计项存在一些偏态。总体而言，图形诊断的结果显示该模型是一个稳定的模型。

对于MCMC诊断（已纳入statnet包）感兴趣的读者，可以参考普鲁玛及其同事的相关论著（Plummer, Best, Cowles & Vines, 2006），另外，对于MCMC诊断的更一般的信息可以参考考尔斯和卡琳的评述（Cowles & Carlin, 1996）。

第 5 节 | 曲线指数族模型

如果不希望通过先验的方式选择 α 值，我们也可以通过模型估计过程获得最佳 α 值。这种通过模型估计选择 α 值，而不是通过先验拟定 α 值的方式，被称为曲线指数族模型(curved exponential family models, CEF)。利用"Command 37"(命令 37)可以使用 CEF 模型尝试对模型进行重新估计。需要注意的是，有些人认为，基于之前我们已经分析过的诸多模型的拟合效果为基础选择 α 的过程，也可以作为估计 CEF 模型 α 值的一种替换方式。

CEF 模型中对于几何统计项 α 的估计值显示在对应的几何统计项之后。本例中，几何加权度对应的最佳 α 值，在表 3.13 显示为 gwdegree.decay，对应的值是 0.838；几何加权边共享伙伴的最佳 α 值则显示为 gwesp.alpha，对应的值是 0.9451；几何加权二元组共享伙伴的最佳 α 值则显示为 gwdsp.alpha，对应的值是 1.822。表 3.13 高亮显示了上述三个统计项所对应的最佳 α 值。

二元依赖性模型与 CEF 模型在协变量的大小上还是存在一些差异的。与二元依赖性模型相比，CEF 模型的 AIC 和 BIC 两项测量结果都降低了，显示模型拟合优度有了提升。我们如果将二元依赖性模型与 CEF 模型均纳入表 3.11 进行

表 3.13　曲线指数族模型统计摘要表

```
==========================
Summary of model fit
==========================
Formula:    lhds ~ edges + nodefactor("years") + nodecov("popmil") + nodematch("hivscreen",
    diff = T, keep = 2) + nodematch("nutrition", diff = T, keep = 2) +
    nodematch("state") + gwdegree(1, F) + gwesp(1, F) + gwdsp(1,
    F)
Iterations:  20
Monte Carlo MLE Results:
                       Estimate Std. Error   MCMC %      Lower         OR    Upper p-value
edges                -9.118e+00 7.742e-01 8.700e+01 2.405e-05 1.097e-04   0.001 < 1e-04 ***
nodefactor.years.1    1.312e-01 4.234e-02 2.000e+00 1.049e+00 1.140e+00   1.239 0.00195  **
nodefactor.years.2    2.402e-01 4.142e-02 7.000e+00 1.172e+00 1.272e+00   1.379 < 1e-04 ***
nodefactor.years.3    2.791e-01 4.081e-02 5.000e+00 1.220e+00 1.322e+00   1.432 < 1e-04 ***
nodecov.popmil        2.253e-01 3.043e-02 2.400e+01 1.180e+00 1.253e+00   1.330 < 1e-04 ***
nodematch.hivscreen.Y 2.079e-01 3.603e-02 1.000e+00 1.147e+00 1.231e+00   1.321 < 1e-04 ***
nodematch.nutrition.Y 1.884e-01 4.348e-02 2.000e+00 1.109e+00 1.207e+00   1.315 < 1e-04 ***
nodematch.state       4.923e+00 9.404e-02 5.000e+00 1.143e+02 1.374e+02 165.187 < 1e-04 ***
gwdegree              1.109e+00 1.841e-01 3.100e+01 2.113e+00 3.031e+00   4.348 < 1e-04 ***
gwdegree.decay        8.380e-01 8.388e-01 8.600e+01 4.467e-01 2.312e+00  11.965 0.31774
gwesp                 9.690e-01 3.487e-02 6.000e+00 2.461e+00 2.635e+00   2.822 < 1e-04 ***
gwesp.alpha           9.451e-01 3.057e-02 0.000e+00 2.424e+00 2.573e+00   2.732 < 1e-04 ***
gwdsp                -8.237e-02 7.573e-02 6.900e+01 7.939e-01 9.209e-01   1.068 0.27677
gwdsp.alpha           1.822e+00 1.196e-01 3.600e+01 4.894e+00 6.187e+00   7.822 < 1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    Null  Deviance: 1140093  on 822403  degrees of freedom
 Residual Deviance:   16987  on 822389  degrees of freedom
          Deviance: 1123106  on     14  degrees of freedom
AIC: 17015     BIC: 17178
```

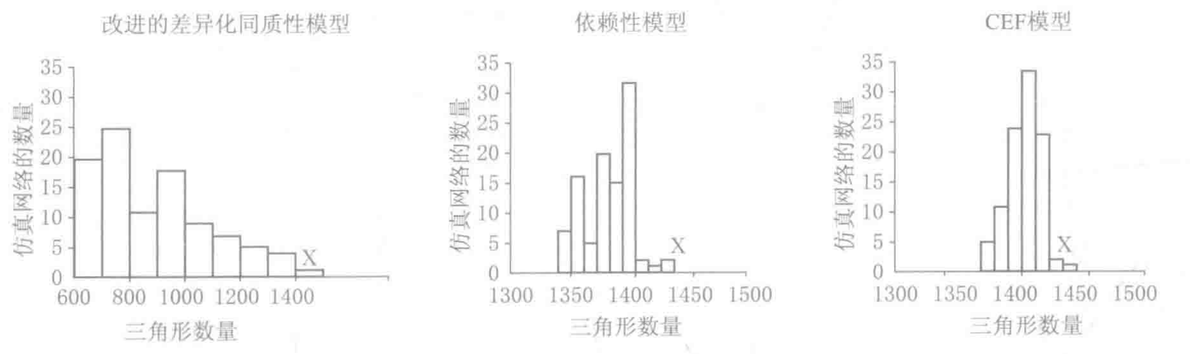
比较,就可以比较具有不同网络结构特征的模型,对于网络拟合优度评价结果的影响(Command 38,参见表 3.14)。

表 3.14 地方卫生机构网络与模型仿真网络在网络拟合测量方面的差异

	edges	degree0	degree1	degree2	degree3	degree4	degree5	triangle
LHD	2708	58	117	182	223	226	172	1437
Null	2647	18	97	159	243	276	196	17
Main effects	2660	29	95	166	243	246	202	32
Homophily	2704	48	127	149	234	244	168	1223
Diff homophily	2707	45	125	169	224	231	174	1249
Diff homophily 2	2713	48	112	182	222	233	170	1249
Dependence	2589	26	129	207	254	207	177	1151
CEF model	2652	54	135	218	195	198	150	1306

网络仿真结果显示:不同模型在网络仿真效果上存在差异,从结果而言,简易的模型包含三角形数量较少,而 CEF 模型的仿真网络在三角形数量方面表现最为出色;另外,三个二元独立同质性模型(同质性模型、差异化同质性模型以及改进的差异化同质性模型)在仿真网络的总边数方面表现得较好,其中,改进的差异化同质性模型在仿真网络的度分布方面表现得最好。然而,上述结果似乎表明并不存在单一的最佳模型能够拟合网络的全部特性,即便是模型仿真效果最佳的模型也并不意味着就能够很好地拟合真实的观测数据。

经过 100 次网络仿真实验后,我们发现差异化同质性模型、依赖性模型以及 CEF 模型均会低估观测网络中的三角形数量,即,通过模型仿真构建的网络所包含的三角形数量均低于观测网络中实际包含三角形数量(该数量在图形中表现为 X)(Command 39,图 3.12)。然而,基于依赖性模型和 CEF 模型进行仿真的网络所包含的三角形数量更接近于观测网络中的三角形数量,其中,CEF 模型是在各方面最为接近观测网络的。在 100 个仿真网络中,依据依赖性模型和



直方图对差异化同质性模型以及两个依赖性模型的仿真效果进行了比较，结果显示经过 100 次模型仿真后，这些模型仿真网络各自对应的整体三角形数量分布。其中，X 代表地方卫生机构网络中实际的三角形数量。

图 3.12

CEF 模型进行仿真获得的三角形数量均没有观测网络中的三角形数量多；仅有 1 个 CEF 模型仿真的结果为 1 437 个三角形；因此，虽然上述两种模型较二元独立性模型的仿真效果更为优异，但两者在对地方卫生机构网络的传递性特征方面并没有表现得特别出色。

另外，对于中心度、距离、边共享伙伴、二元组共享伙伴等特征的可视化展现，也显示出依赖性模型与 CEF 模型在提升网络拟合优度方面的差异，参见"Command 40"（命令 40），图 3.13。上述两个模型在网络仿真的效果方面显示出与地方卫生机构网络很大的相似性，不过，CEF 模型在中心度特征的把握上更接近地方卫生机构网络。

模型选择

通过比较前述七种模型（零模型、主效应模型、同质性模型、差异化同质性模型、改进的差异化同质性模型、依赖性模型以及 CEF 模型）的多个统计和图形化拟合优度测量指标，我们发现：CEF 模型有最佳的网络仿真效果。在模型构建过程中，对于网络仿真效果提升影响最大的地方源于之前模型的两个改进：增加同质性统计项，即考虑具有相似特征（州、项目）的地方卫生机构之间建立沟通关系的影响；以及增加依赖性统计项，即考虑度分布特征以及传递性特征对地方卫生机构之间建立沟通关系的影响。当然，除了对上述模型进行整体的拟合优度测量以及仿真效果的图形化展现的方法之外，对于较小的网络可以对网络构建过程中不同阶段的模型拟合度进行测量。这种分解的方法十分有利。对于像地

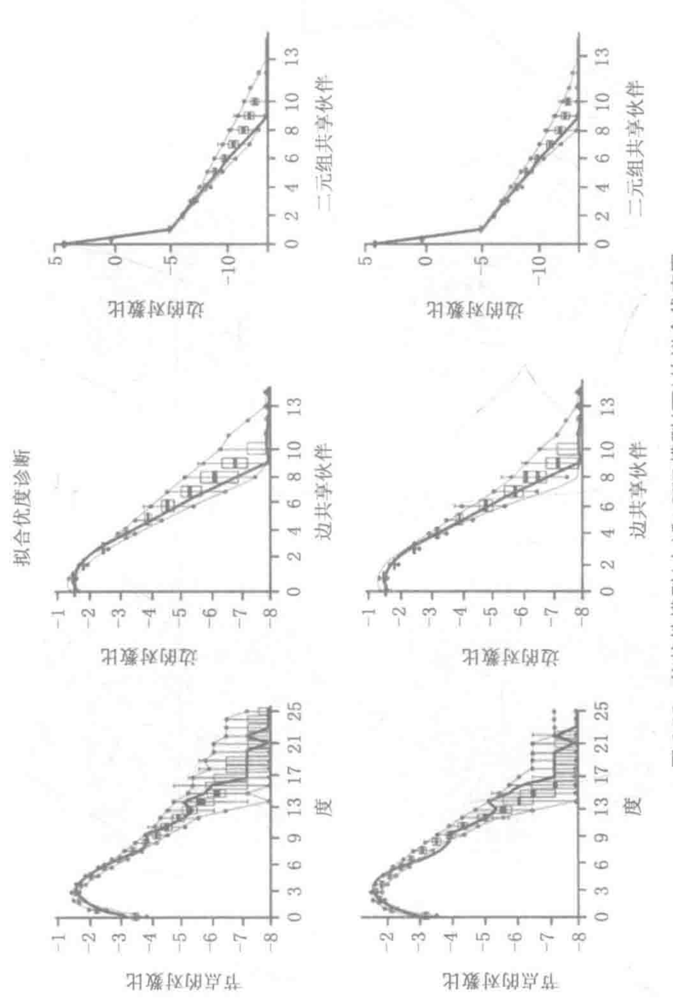


图 3.13　依赖性模型(上)和 CEF 模型(下)的拟合优度图

方卫生机构网络这样规模较大的网络而言，想通过可视化观察方法来识别出同质性统计项对仿真效果的影响是十分困难的(Command 41，图 3.14)。

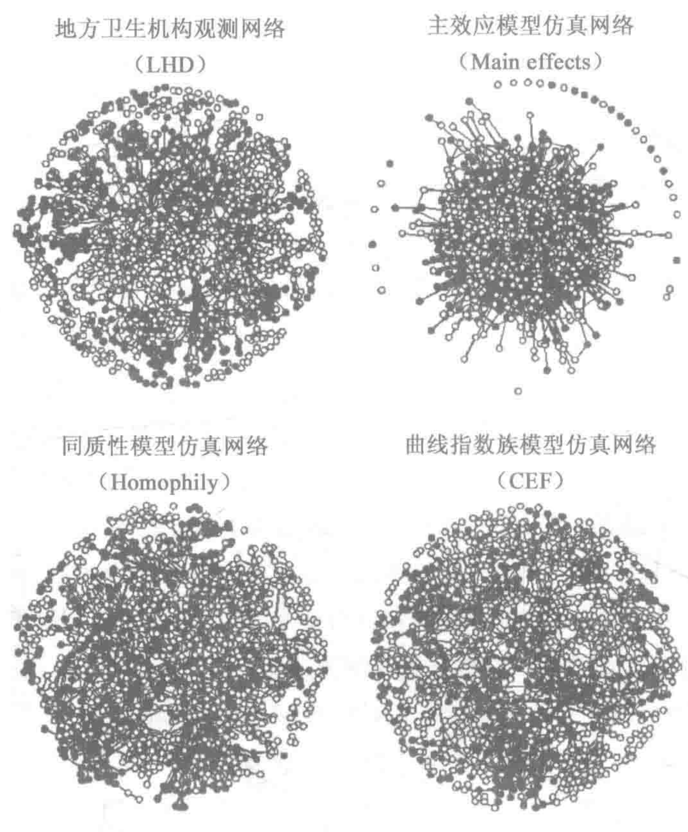


图 3.14 地方卫生机构观测网络以及三种模型的仿真网络

虽然通过同质性模型和 CEF 模型获得的仿真网络在结果上十分相似，但值得注意的是，CEF 模型在执行项目机构之间的聚类模式上表现得与观测的地方卫生机构网络极为

相似。由于CEF模型在AIC和BIC指标以及仿真网络拟合优度方面表现出了优势，因此，CEF模型就更有可能被选择作为最终采纳的模型。如图3.14那样的一张图或者如表3.15那样的一张表，就能够展示模型的构建过程，这种方法不仅可以用来阐述模型的发展历程，也可以作为最终模型选定的理由。

依赖性模型的解释

对依赖性模型的同质性统计项进行检验的结果是显著的，如执行艾滋病筛查项目、执行营养项目以及同处一个辖区等。两个均执行了艾滋病筛查项目的地方卫生机构之间建立沟通关系的概率，是网络中其他机构之间建立沟通关系的概率的1.23倍。另外，地处同一个州对于沟通关系的建立也存在显著的影响（OR＝137.4；95％CI＝114.3－165.2）。同样，两个均执行了营养项目的地方卫生机构之间对于沟通关系建立概率的影响也是显著的，两个执行了营养项目的地方卫生机构之间建立沟通关系的概率是两个均没有执行营养项目的机构之间建立沟通关系概率的1.21倍（95％CI＝1.11－1.32）。表3.15概括了四个模型的发展历程，展示了模型系数的估计值以及标准差；根据用户需求的情况，上述表还可以包含优势比及置信区间等参数。

一旦GW（几何加权）统计项被增加到模型中来，预测任意两个网络成员之间关系形成的概率就变得复杂，原因在于：对几何加权项的变化统计进行计算和解释比较困难。以几何加权度项为例，网络中每增加一条边，$D_i(y)$和$D_{i+1}(y)$

表 3.15　零模型、主效应模型、改进的差异化同质性模型以及 CEF 模型的统计摘要表(部分)

	Estimate(SE)			
	Null Model	*Main Effects*	*Differential Homophily 2*	*CEF*
Edges(constant)	−5.71(0.02)	−6.23(0.06)	−9.56(0.11)	−9.12(0.77)
Main effects				
Population(millions)		0.20(0.01)	0.33(0.02)	0.23(0.03)
Years experience				
1—2		Reference	Reference	Reference
3—5		0.14(0.05)	0.18(0.05)	0.13(0.04)
6—10		0.28(0.04)	0.32(0.04)	0.24(0.04)
11+		0.34(0.04)	0.35(0.04)	0.28(0.04)
Homophily				
State			6.31(0.08)	4.92(0.09)
Conducts nutrition program			0.25(0.05)	0.19(0.04)
Conducts HIV screening			0.46(0.04)	0.21(0.04)
Structural terms				
GWD				1.11(0.18)
GWESP				0.97(0.03)
GWDSP				−0.08(0.08)
Fit				
AIC	36367	36176	19473	17015
BIC	36379	36234	19566	17178

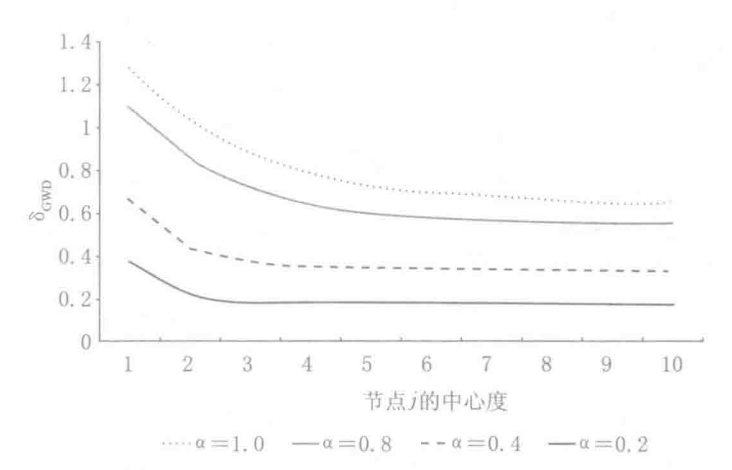
项作为公式 2.9 右侧加权度求和的部分，将会被 $D_i(y)-1$ 和 $D_{i+1}(y)+1$ 项取代。当网络其他因素保持不变时，为了检验网络中每增加一条边对于网络图形形成概率的影响，我们需要将网络的新旧度值(增加一条边前后网络的度值)代入到公式 2.8 中来，计算网络中图的优势比(Hunter, 2007)。读者如果有兴趣进一步了解如何通过代入法计算模型几何加权度的变化统计结果，也可以就下列问题参考亨特的论著(Hunter, 2007)：

$$\frac{P(Y_{ij}=1)_{after}}{P(Y_{ij}=1)_{before}}=\exp\{\theta(1-e^{-\alpha})^{i}\} \qquad [3.5]$$

需要注意的是，$P(Y_{ij}=1)_{before}$ 是 $P(Y_{ij}=1|n \text{ actors}, Y_{ij}^c)$ 的简写形式。由于网络新增了一条边，于是，与该边相连的两个节点的中心度都会有所增加，增加转而变成了 $\theta[(1-e^{-\alpha})^i+(1-e^{-\alpha})^j]$(Hunter, 2007)。因此，几何加权度的变化统计就可以表示为：

$$\delta_{GWD}=(1-e^{-\alpha})^i+(1-e^{-\alpha})^j \quad [3.6]$$

值得注意的是，当中心度增加时，$(1-e^{-\alpha})^i$ 呈几何级下降。所以，如果 θ_{GWD} 是正向且显著的，那么，对于所有 i 和 j 中心度值而言，边的对数优势比也会增加；不过，当 i 和 j 中心度值已经很高时，对数优势比的增加速率就会下降；当到达某一个临界点时，即 j 已经达到一个高中心度的阶段后，对数优势比的增长也会趋于平稳并保持一个常数。对于一个较小的 α 而言，网络常常会更快就出现这种平衡趋势（参



当 α 的取值不同时，节点 i（度为1）与不同中心度值的节点 j 之间建立联系的对数优势比的变化情况。

图 3.15

见图 3.15)。GWD 网络的统计结果对于中心度赋予的权重越高,网络中具有高中心度的节点的统计值就越大。变换统计的趋势表明,在具有低中心度的节点之间,增加一条连线的倾向是最强的。

几何加权边共享伙伴(GWESP)与几何加权二元组共享伙伴(GWDSP)对于整体网络结构的变化的影响更为复杂。因为,网络中每增加一条边不仅会改变边共享伙伴(ESP)也会改变二元组共享伙伴(DSP)的数量,不仅涉及相关节点而且包括与这两个节点相关的网络中全部的节点。对变化统计值受 GWESP 和 GWDSP 何种影响感兴趣的读者,可以参考亨特的论著(Hunter, 2007):

$$\delta_{GWESP} = (1 - e^{-\alpha})^{ij_{ESP}} \tag{3.7}$$

$$\delta_{GWDSP} = (1 - e^{-\alpha})^{ij_{DSP}} \tag{3.8}$$

在地方卫生机构依赖性模型中,当 $\alpha = 1.0$ 时,对应的计算结果是:

$$(1 - e^{-\alpha}) = 1 - e^{-1.0} = 0.63$$

可以(将上述结果)代入公式 3.6 和公式 3.8,计算三个条件项的变化统计值:

$$\delta_{GWD} = 0.63^{i_d} + 0.63^{j_d}$$

$$\delta_{GWESP} = 0.63^{ij_{ESP}}$$

$$\delta_{GWDSP} = 0.63^{ij_{DSP}}$$

δ_{GWD}显示了两个节点 i 和 j 之间建立一项连接,且这两个节点的中心度均为 0 时的对数优势比增加得最多;随着两个节点 i 和 j 的联系数量增多,联系之间的对数优势比的增

加幅度就会大幅下降（Hunter，2007）。同样地，假定网络中其他节点之间的关系保持不变，两个特定网络成员之间建立联系的概率也可以使用 GWESP 和 GWDSP 的变换统计来计算。随着 GWD、GWDSP 和 GWESP 的变化统计值的增加，网络的对数优势比呈现下降趋势的现象，可以被称为“异配倾向”（antipreferential attachment）（Hunter，2007）。

一般而言，对 GWD、GWDSP 和 GWESP 系数的解释和对其他模型系数的解释是一样的。一个几何统计项具有正向且显著的系数可以解释为：当网络中其他因素保持不变时，两个任意给定节点 i 和 j 之间建立联系的概率将比这两点之间随机发生联系的概率要大；同样地，一个负向且显著的系数可以解释：两个任意给定节点 i 和 j 之间建立联系的概率将比这两点之间随机发生联系的概率要小；一个非显著性的系数则可以解释为：当其他因素保持不变时，在节点 i 和 j 之间建立联系的概率较变换之前并没有显著的变化。

虽然这些系数看似很简单，但变化统计结果却能反映更深层次的含义。正如前面所描述的，变化统计目的在于：通过观察节点 i 和 j 之间增加一条连线对（Hunter & Goodreau et al.，2008）整个网络统计值变化的影响情况（参见第 2 章公式 2.9 至公式 2.11）。考虑到网络中增加一条边对于整个网络的共享伙伴分布的影响，我们在对 GWESP 和 GWDSP 系数进行解释时就需要格外注意，避免过度解释这两个系数。根据亨特的方法（Hunter，2007：第 5 部分），应当明确对于 GWESP 和 GWDSP 系数的解释的基础在于“假定网络没有其他变化因素需要考虑，以及所有其他的模型效果均已考虑”（Hunter，2007：227）。

最终，由于 DSP 测量的是连接了或未连接的二元组共享伙伴的情况，而 ESP 则仅测量那些连接了的二元组共享伙伴的情况，因此，一个值得重点关注的地方是：我们需要单独看这些增加到模型中的统计项，也要将他们合起来看。如果模型中仅包含 GWDSP 而不包括 GWESP，那么，最终的系数值会同时受到相连以及不相连的共享伙伴分布的影响。如果模型中仅包含 GWESP 而不包括 GWDSP，那么，最终的系数值仅受到相互连接的二元组共享伙伴分布情况的影响。如果 GWESP 和 GWDSP 同时被考虑，GWESP 仍是考虑或控制连接二元组共享伙伴分布情况，但允许 GWDSP 考虑未连接的二元组之间的共享伙伴分布。

模型中增添了 GW（几何加权）统计项，也就增加了需要利用模型进行预测的信息量。例如，如同之前情形，对两个地方卫生机构建立沟通关系的概率进行预测时，一个地方卫生机构拥有 1 年履职经验的领导（years＝0），10 万辖区选民（popmil＝0.1），没有执行过艾滋病筛查项目（hivscreen＝0），但却执行了营养项目（nutrition＝1）；另一个地方卫生机构拥有 7 年履职经验的领导（years＝2），200 万辖区选民（popmil＝2），执行过艾滋病筛查项目（hivscreen＝1），也执行过营养项目（nutrition＝1）。利用主效应模型预测这两个地方卫生机构之间建立沟通的概率为 0.0023，而利用差异化同质性模型预测这两个地方卫生机构之间建立沟通关系的概率为 0.033。想要采用依赖性模型预测两个地方卫生机构建立沟通关系的概率时，我们不仅仅需要知道两个网络成员的属性特征，还需要知道二元组中每一个节点所各自具有的中心度、边共享伙伴数量以及二元组共享伙伴数量。此前已经获

得的GWD、GWESP和GWDSP的变化统计值将会与网络节点属性的系数及变化统计值一并纳入到模型中来：

$$\begin{aligned}P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(&-10.07 * \delta_{\text{edges}} \\ &+ 0.20 * \delta_{\text{popmil}} + 0.14 * \delta_{3-5\text{years}} + 0.25 * \delta_{6-10\text{year}} \\ &+ 0.30 * \delta_{>10\text{years}} + 0.19 * \delta_{\text{HIVHom}} + 0.18 * \delta_{\text{nutritionHom}} \\ &+ 5.02 * \delta_{\text{stateHom}} + 0.19 * \delta_{\text{GWD}} + 0.96 * \delta_{\text{GWESP}} \\ &- 0.04 * \delta_{\text{GWDSP}})\end{aligned}$$

上述计算过程获得的模型预测值可以代入模型，预测某些特定情形下，网络新增一条边的概率。由于模型已经包含了众多的统计项，因此，在下面的计算中，我们仅选择显示那些在计算概率过程中发挥作用的统计项。案例1重新审视了两个地方卫生机构建立沟通关系的概率，使用主效应模型的预测结果为0.0023或者0.23%，而使用差异化同质性模型的预测结果则为0.033或者3.3%。由于依赖性模型包括了度、边共享伙伴和二元组共享伙伴等结构性统计项，因此，我们必须在计算过程中提供两个地方卫生机构的上述结构性统计值。

案例1：一个位于密苏里州的地方卫生机构，其领导具有1年的履职经验，该机构所属辖区拥有10万选民，该机构没有执行过艾滋病筛查项目，但执行过营养项目；另一个位于密苏里州的地方卫生机构，其领导具有7年的履职经验，该机构所属辖区拥有200万选民，该机构执行过艾滋病筛查项目和营养项目；这两个机构各自的中心度为3和4，以及0个边共享伙伴和3个二元组共享伙伴。计算过程如下：

$$\begin{aligned}P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(&-10.07 * 1 \\ &+ 0.20 * 2.1 + 0.25 * 1 + 0.18 * 1 + 5.02 * 1\end{aligned}$$

$$+0.19*(0.63^3+0.63^4)+0.96*0.63^0-0.04*0.63^3)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=\text{logistic}(-3.17)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=0.040$$

案例 2:两个位于俄勒冈州的地方卫生机构的领导均具有 10 年履职经验,两家机构所属的辖区均有 2.5 万选民,均执行了艾滋病筛查项目和营养项目,其各自对应的中心度为 2 和 4,具有 1 个边共享伙伴和 2 个二元组共享伙伴。计算过程如下:

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=\text{logistic}(-10.07*1+0.20*0.05$$
$$+0.30*2+0.19*1+0.18*1+5.02*1$$
$$+0.19*(0.63^2+0.63^4)+0.96*0.63^1$$
$$-0.04*0.63^2)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=\text{logistic}(-3.38)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=0.033$$

案例 3:两个位于加利福尼亚州的地方卫生机构的领导均具有 10 年履职经验,两个机构所属的辖区均有 200 万选民,均执行了艾滋病筛查项目和营养项目,其所对应的中心度为 2 和 4,另包含 1 个边共享伙伴和 2 个二元组共享伙伴。计算过程如下:

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=\text{logistic}(-10.07*1+0.20*4$$
$$+0.30*2+0.19*1+0.18*1+5.02*1$$
$$+0.19*(0.63^2+0.63^4)+0.96*0.63^1-0.04*0.63^2)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=\text{logistic}(-2.59)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c)=0.070$$

这些案例的计算结果给我们提供了观察地方卫生机构

建立沟通关系网络的新视角。例如，辖区人口规模的大小因素并不是孤立存在的；相反，当它与机构二元组中两个机构辖区人口相乘时，它能使沟通关系产生的概率发生巨大变化；案例2和案例3仅仅是调整了两个地方卫生机构辖区人口规模，但网络中建立沟通关系的概率从3.3%急速增长到7.0%。

预测CEF模型的概率更加复杂，因为每一个几何统计项的估计是以α值为基础的。本例中，CEF模型首先计算出了一系列α值，对于GWD而言，α值为0.838，对于GWESP而言，α值为0.9451，而对于GWDSP而言，α值为1.822。因此，我们可以利用这些α值，进而计算每一个几何统计项的基准值，并以此为基础最终计算出每一个几何统计项的变化统计值：

$$(1-e^{-\alpha})=1-e^{-0.838}=0.57$$

$$(1-e^{-\alpha})=1-e^{-0.9451}=0.61$$

$$(1-e^{-\alpha})=1-e^{-1.822}=0.84$$

每一个基准值都可以代入对应公式来决定三个几何统计项的变化统计值：

$$\delta_{GWD}=0.57^{i_d}+0.57^{j_d}$$

$$\delta_{GWESP}=0.61^{ij_{esp}}$$

$$\delta_{GWDSP}=0.84^{ij_{dsp}}$$

于是，完整的CEF模型和系数就可以表示为：

$$\begin{aligned}P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c})=\text{logistic}(&-9.12*\delta_{edges}\\&+0.23*\delta_{popmil}+0.13*\delta_{3-5years}+0.24*\delta_{6-10years}\\&+0.28*\delta_{>10years}+0.21*\delta_{HIVscreenHom}+0.19*\delta_{nutritionHom}\\&+4.92*\delta_{stateHom}+0.11*\delta_{GWD}+0.97*\delta_{GWESP}\\&-0.8*\delta_{GWDSP})\end{aligned}$$

于是，上面的三个案例所预测的概率也会被计算出来：

案例 1：一个位于密苏里州的地方卫生机构，其领导具有 1 年的履职经验，该机构所属辖区拥有 10 万选民，该机构没有执行过艾滋病筛查项目，但执行过营养项目；另一个位于密苏里州的地方卫生机构，其领导具有 7 年的履职经验，该机构所属辖区拥有 200 万选民，该机构执行过艾滋病筛查项目和营养项目；这两个机构各自的中心度为 3 和 4，以及 0 个边共享伙伴和 3 个二元组共享伙伴。计算过程如下：

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(-9.12*1+0.23*2.1 + 0.24*1+0.19*1+4.92*1+0.11*(0.57^{3} + 0.57^{4})+0.97*0.61^{0}-0.8*0.84^{3})$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(-2.33)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = 0.088$$

案例 2：两个位于俄勒冈州的地方卫生机构，两个机构的领导均具有 10 年履职经验，两家机构所属的辖区均具有 2.5 万选民，均执行了艾滋病筛查项目和营养项目，其各自对应的中心度为 2 和 4，具有 1 个边共享伙伴和 2 个二元组共享伙伴。计算过程如下：

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(-9.12*1+0.23*0.05 + 0.28*2+0.21*1+0.19*1+4.92*1 + 0.11*(0.57^{2}+0.57^{4})+0.97*0.61^{1} - 0.08*0.84^{2})$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = \text{logistic}(-2.65)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^{c}) = 0.066$$

案例 3：两个位于加利福尼亚州的地方卫生机构，两个机

构的领导均具有10年履职经验，两家机构所属的辖区均具有200万选民，均执行了艾滋病筛查项目和营养项目，其所对应的中心度为2和4，另包含1个边共享伙伴和2个二元组共享伙伴。计算过程如下：

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c) = \text{logistic}(-9.12*1+0.23*4 +0.28*2+0.21*1+0.19*1+4.92*1 +0.11*(0.57^2+0.57^4)+0.97*0.61^1-0.08*0.84^2)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c) = \text{logistic}(-1.74)$$

$$P(Y_{ij}=1 \mid n \text{ actors}, Y_{ij}^c) = 0.150$$

用约束条件重新定义模型

到目前为止，模型建构中关注的主要问题是：哪些类型的统计项应该纳入到模型中来，以及模型估计过程中有哪些推荐的设定条件。另外，在解决模型建构过程问题时，如果能对可能产生的仿真网络范围进行限定也将有利于某些研究问题。例如，在网络数据调研中，限定可能被提名人员的数量等。在这种情况下，限定仿真网络中节点的最大中心度将会是有效的做法。这些限制条件既可以是限定度的最大值或者最小值，也可以是限定节点具有同样的中心度，还可以是限定节点符合某个度分布，抑或是仅限定网络的边数。对于这些限定条件更多的细节可以参考莫瑞斯和他的同事们(Morris et al., 2008)的资料，也可以通过在R提示符中后输入R-ergm帮助文档获得这些限定的定义。由于我们对于任何给定网络潜在发生作用的社会力量知之甚少，因此，往往不建议在获取观察网络的方式上采取限定措施。然而如果有必要，这个函数是可获得的。



第4章

面向有向网络及二元组属性的应用

第1节 针对有向网络的研究

与之前介绍的无向网络相似，有向的观测网络在度分布以及传递三角形数量等特征上经常是不同于随机网络的。除此之外，莫雷诺在他1934年的著作中也提到过有向的观测网络更易于产生互惠关系(Moreno，1934，1953)。为了验证这些网络特征，我们需要建立一个随机的有向网络，并利用该随机网络和波莫纳湖(Lake Pomona)网络进行比较，关于波莫纳湖的网络数据已经包含在R的网络包中。正是由于该数据已经包含在了R的数据包中，因此，我们可以很方便地将波莫纳湖网络数据导入R(Command 42)。波莫纳湖网络数据展现了搜救行动中组织间的交互关系，更多关于该数据的信息可以从该数据包中获得(http://cran.r-project.org/web/packages/network/network.pdf)。

在波莫纳湖网络中，从节点A至节点B(A→B)之间的一条连线表明A组织在其报告中提及与B组织之间存在交互关系。整个网络包括了20个组织以及组织间的148条有向链接，网络密度为0.39。根据对图4.1的初步观察，波莫纳湖网络(左侧)和与其具有同样规模和密度的随机网络(右侧)之间存在结构上的差异，这种差异酷似我们讨论的无向网络及其随机网络之间存在的差异，即观测网络中心由一群紧密

连接的组织群体构成，观测网络的外围则由一群稀疏连接的组织群体构成，而随机网络图中群体之间链接的分布似乎更加均匀(Command 43)。

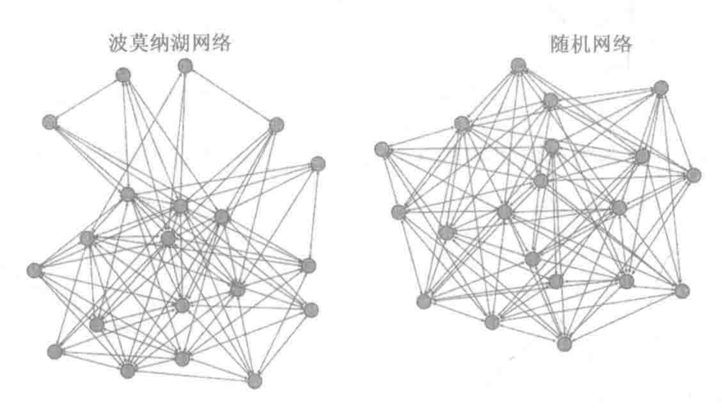


图 4.1　波莫纳湖网络与随机网络的图形比较

当进一步对比观测网络与随机网络的入度与出度分布时发现，波莫纳湖网络有一个右倾的入度分布模式，这一点十分类似于地方卫生机构(LHD)网络的度分布特征(参见图 4.2)，而随机网路与期望的情况一致，入度分布表现为更加均匀的入度分布模式；波莫纳湖网络的出度分布则没有显示出十分清楚的模式，仅有一点右倾，而随机网络则是接近于均匀分布(Command 44)。

有向网络反映出的另一个特征是交互性(例如 A↔B)和非对称(例如 A→B，A←B)关系同时存在。我们可以利用二元组测量方法对波莫纳湖网络和随机网络在上述结构特征进行数量上的比较，从而揭示这种网络结构特征(Command 45，结果参见图 4.3)。

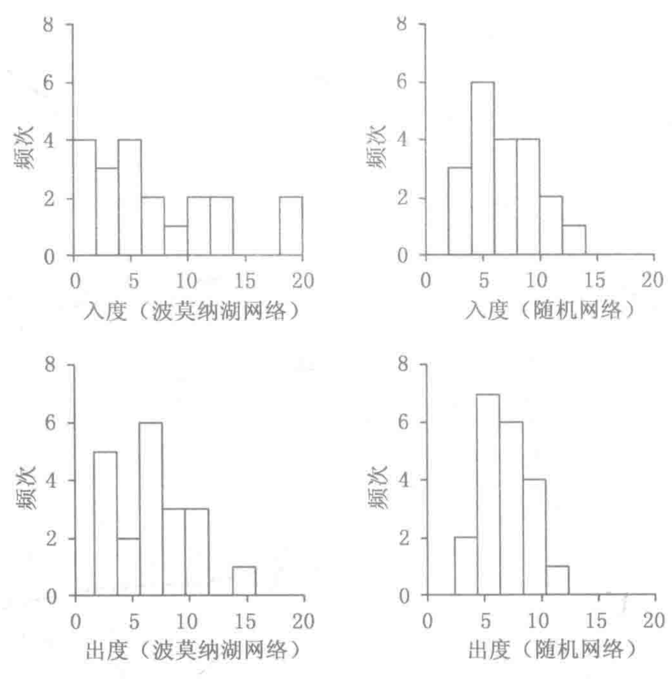


图 4.2 波莫纳湖网络(左侧)和随机网络(右侧)的度分布比较

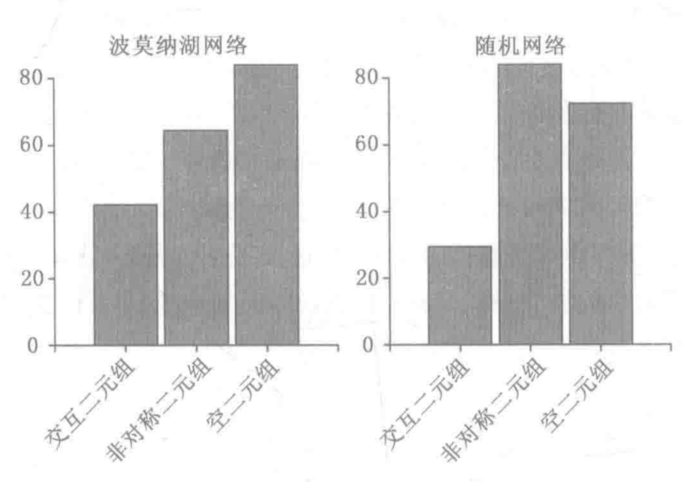


图 4.3 波莫纳湖网络(左)和随机网络(右)的二元组类型分布比较

与无向网络相比，有向网络存在更多的三元组类型（参见附录 C）。因此，波莫纳湖网络和随机网络在三元组的各种类型所对应的频次上也存在差异。波莫纳湖网络在 003 类和 300 类的三元组上频次更高（Command 45；参见附录 C）。这两个类别分别代表三元组中三个节点完全不相连或者完全相连的状态。出现较高频次的 003 类和 300 类三元组现象，与之前在波莫纳湖网络二元组测试中体现出的高频次的交互二元组与空二元组情形类似（参见图 4.3），因为一个 300 类三元组需要由三个交互二元组构成，同时，一个 003 类三元组也需要由三个空二元组构成。

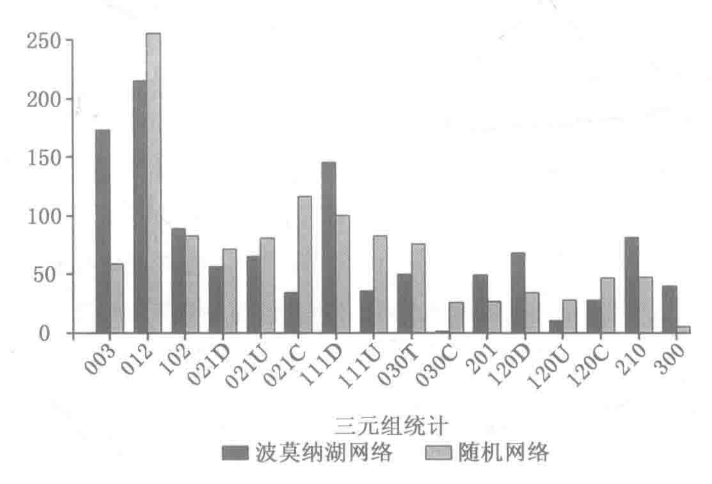


图 4.4　波莫纳湖网络（左）和随机网络（右）的三元组分布比较

p_1 模型

尽管 p_1 模型在指数随机图建模（ERGM）的发展历程中具有突出的地位（参见第 2 章），但在此前构建地方卫生机构

指数随机图模型的过程中，我们没有采用 p_1 模型，这是因为 p_1 模型是专门针对有向网络结构特征设计的，因此，它并不适合类似地方卫生机构这样的无向网络数据，但却适用于波莫纳湖网络。

波莫纳湖网络的零模型统计结果显示(Command 47)，零模型的边统计项是显著的(b=−0.45; SE=0.11)，对应零模型的赤池信息准则(AIC)和贝叶斯信息准则(BIC)结果分别为 510.1 和 514.0。霍兰德和莱因哈特(Holland & Leinhardt, 1981)的 p_1 模型整合了四种二元组效应：(1)边的数量或考虑边数量的网络密度；(2)发送者效应(sender terms)，或者说是发出链出关系的节点属性(扩展性)；(3)接受者效应(receiver terms)，或者说是接受链入关系的节点属性(吸引性)；(4)交互效应(互惠性)。尽管交互统计项将二元组视为彼此独立，但 AB 节点对与 BA 节点对之间本质上是相互依赖的，因此，就可以考虑应用马尔科夫链蒙特卡洛(MCMC)估计方法(Morris et al., 2008)。发送者和接受者效应作为 p_1 模型构建的基础，实际上任何一个节点对既包括一条发送者的链出关系也包括了一条接受者的链入关系，因此，实际测量中应去掉一种关系从而避免信息出现重复的问题。在这种情况下，模型将包括边统计项、19 个对应的发送者统计项、19 个接受者统计项以及一个交互统计项，共计 40 项统计项。需要注意的是，运行命令 47 的结果中包括一些统计项，这些统计项对应的系数值为 Inf 或−Inf。如果 Inf 和−Inf 值出现在指数随机图模型中，表明该系数为极大值(Inf)或者极小值(−Inf)。这种情况下，很有可能说明这些组织接受了很少或者没有接收到联系。估计的结果中包含这些极值通常

会导致模型无法估计离差和相关统计结果，于是，模型拟合结果将无法与零模型进行比较。然而，通过设定参数的方法（类似于第 3 章中所介绍的在“nodematch”命令中设定 keep 参数的方法），就可以除去 Inf 问题的参数，模型将会依据最适合的统计项进行重新估计（Command 48，参见表 4.1）。

表 4.1　波莫纳湖网络搜救行动中的 p_1 模型

```
==========================
Summary of model fit
==========================

Formula:   lake ~ edges + sender + receiver(base = c(1, 5, 8, 19))
 + mutual

Iterations:   20

Monte Carlo MLE Results:
           Estimate Std. Error MCMC %   p-value
edges        0.2064     0.5950      0  0.728963
sender2    - 0.2751     0.7544      0  0.715621
sender3    - 1.2372     0.7678      0  0.108056
sender4    - 0.4869     0.7573      0  0.520725
sender5      0.2770     0.7451      0  0.710275
sender6    - 0.8070     0.7544      0  0.285504
sender7    - 1.0905     0.7552      0  0.149686
sender8    - 2.0586     0.8142      0  0.011908 *
sender9    - 1.2386     0.7662      0  0.106896
sender10   - 2.6285     0.9610      0  0.006557 **
sender11   - 0.9303     0.7681      0  0.226636
sender12   - 1.7683     0.8243      0  0.032645 *
sender13   - 2.6256     0.9651      0  0.006850 **
sender14     0.2303     0.7601      0  0.762054
sender15     2.0868     0.8727      0  0.017335 *
sender16   - 1.1603     0.7766      0  0.136062
sender17   - 0.5686     0.7552      0  0.452031
sender18   - 1.4730     0.8222      0  0.074078 .
sender19   - 1.0969     0.7507      0  0.144864
sender20   - 1.3131     0.7928      0  0.098583 .
receiver2  - 1.0747     0.6300      0  0.088951 .
```

续表

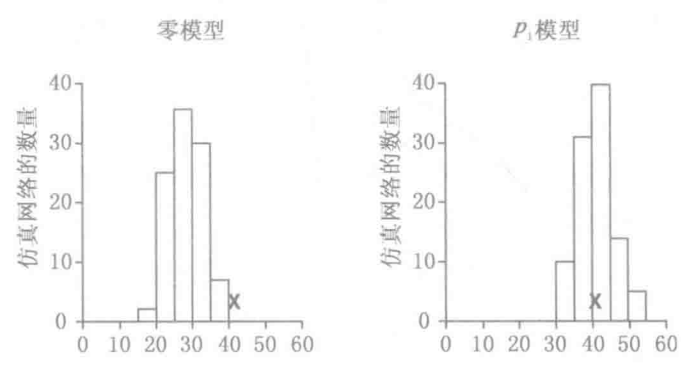
```
receiver3      0.5489   0.6054    0 0.365203
receiver4     -1.3387   0.6522    0 0.040867 *
receiver6      1.0067   0.6324    0 0.112336
receiver7      1.0823   0.6305    0 0.086957 .
receiver9      0.5477   0.6042    0 0.365314
receiver10    -1.1628   0.6943    0 0.094881 .
receiver11    -0.6130   0.6137    0 0.318571
receiver12    -0.9995   0.6517    0 0.126026
receiver13    -1.1658   0.6937    0 0.093757 .
receiver14    -0.9008   0.6108    0 0.141239
receiver15    -2.9824   0.8390    0 0.000432 ***
receiver16    -0.8459   0.6278    0 0.178724
receiver17     0.1097   0.5984    0 0.854678
receiver18    -2.3766   0.8614    0 0.006111 **
receiver20    -1.4664   0.6893    0 0.034094 *
mutual         1.5960   0.4144    0 0.000140 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 526.79  on 380  degrees of freedom
Residual  Deviance: 375.58  on 343  degrees of freedom
          Deviance: 151.21  on  37  degrees of freedom

AIC: 449.58    BIC: 595.37
```

p_1 模型的结果显示 AIC 值要小于零模型，而 p_1 模型 BIC 却上升了。因此该测量结果并没有明确回答 p_1 模型是否优于零模型。与第 3 章的思路一样，对于零模型和 p_1 模型分别进行模型仿真有助于我们对模型拟合优度进行评价(Command 49;图 4.5)。

图中的 X 表示观测网络中交互关系数量所处的位置；p_1 模型较零模型在获取网络的互惠性特征方面表现得更好一些。然而，在解释网络结构方面，p_1 模型忽略了较发送者效应与接受者效应更为有用的节点属性特征。波莫纳湖网



基于零模型(左侧)和 p_1 模型(右侧)进行 100 次模型仿真之后网络中二元组所包含的交互关系数量。其中，X 表示观测网络中二元组所包含的交互关系数量($n=40$)。

图 4.5

络数据也包含组织的一些特征，包括某组织是属于本地组织还是外地组织，组织中雇员的人数，组织隶属情况(是属于城市、州、联邦还是私人机构)以及在各时间点上组织中志愿者的人数。图 4.6 分别用黑色与白色显示了波莫纳湖网络中的

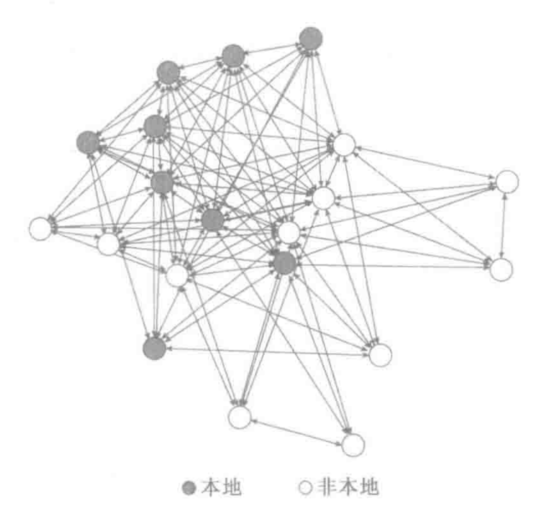


图 4.6　波莫纳湖搜救网络

本地组织和非本地组织(Command 50)。其中用颜色区分了本地组织和非本地组织。

从图 4.6 中可以很明显观察到,在本地组织和非本地组织之间存在一定程度的分离趋势。本地组织之间高度交互地群聚在一起;而非本地组织则处于网络的外围,似乎既与本地组织交流也与非本地组织交流。混合矩阵方法或许有助于厘清这种模式(参见表 4.2)。

表 4.2 针对波莫纳湖搜救网络中机构所处地理位置的混合矩阵

```
        To
From      L NL Total
  L      51 21    72
  NL     40 36    76
  Total  91 57   148
```

注:L 代表本地组织,NL 代表非本地组织。

值得注意的是,在混淆矩阵中的"From"与"To"是有特指的,由于网络之间的链接是有向的,因此,链接表示为从一个节点到另外一个节点的一条弧。在波莫纳湖网络的 148 条链接中,51 条链接是从本地组织指向本地组织的,36 条链接是从非本地组织指向非本地组织的,40 条链接是从非本地组织指向本地组织,还有 21 条链接是从本地组织指向非本地组织。一般而言,无论是本地组织还是非本地组织,似乎都更易于与本地组织进行交流,本地组织较外地组织获得了更多的链入联系(91 vs. 57)。正是由于网络是有向的,因此,许多模型中所包括的统计项就必须进一步明确区分究竟是链入关系(入度)还是链出关系(出度)。例如,在波莫纳湖网络中,本地组织具有很高的入度,于是,我们可以建立一个考虑组织属地特征以及入度特征的主效应模型(Command

51)，通过该模型就可以发现：本地组织比非本地组织在接收到链入关系的概率上高出67%(转移概率[OR]=0.33；95%的置信区间[CI]=0.22—0.51)。表4.3显示了一个主效应模型，该模型包括了最常见边统计项和交互统计项，同时交互统计项伴随着一个nodeifactor参数。这个nodeifactor参数的使用十分类似之前提到的nodefactor参数，但该参数专门用于链入关系；而针对链出关系时，则使用nodefactor参数。

表4.3 在波莫纳湖网络中的主效应模型

```
==========================
Summary of model fit
==========================

Formula:  lake ~ edges + nodeifactor("Location")

Iterations:  20

Monte Carlo MLE Results:
                          Estimate Std. Error MCMC %   Lower     OR Upper p-value
edges                       0.1288    0.1533     NA   0.8424  1.1375 1.536   0.401
nodeifactor.Location.NL - 1.1097    0.2182     NA   0.2150  0.3297 0.506   <1e-04 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ''  1

    Null  Deviance: 526.792  on 380  degrees of freedom
 Residual Deviance: 481.276  on 378  degrees of freedom
          Deviance:  45.515  on   2  degrees of freedom

AIC: 485.28     BIC: 493.16
```

主效应模型的拟合优度图显示：该模型对观测网络的入度、出度及二元组共享伙伴(DSP)特征均进行了很好的拟合。然而，该模型没有能够很好拟合边共享伙伴(ESP)的网络特征(Command 51；图4.7)。为了更好地拟合共享伙伴的网络分布特征，添加几何加权条件可能是种较为有益的

尝试。

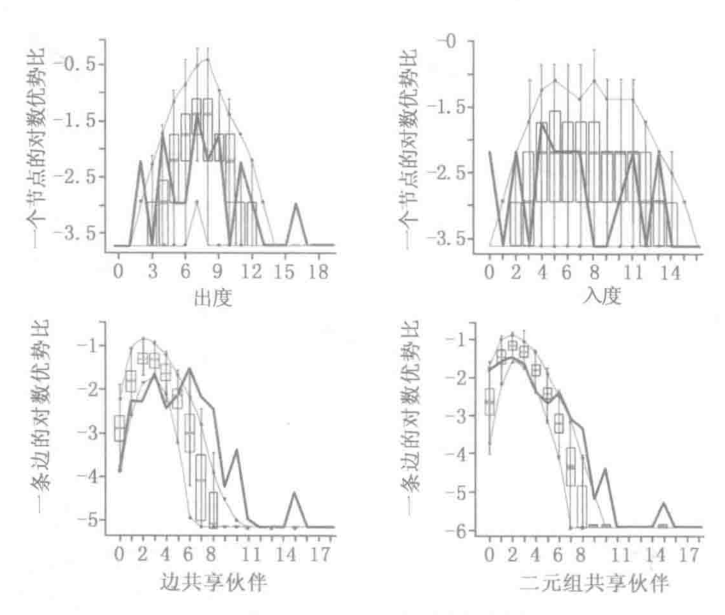


图 4.7 波莫纳湖网络主效应模型的拟合优度图

波莫纳湖流域应急响应协作网络的依赖性模型(表 4.4)包括了几何加权边共享伙伴(GWESP)和几何加权二元组共享伙伴(GWDSP)等统计项,从离差上看($\chi^2(2)=19.8$; $p<0.05$),依赖性模型较主效应模型具有更好的拟合优度,但从图形的拟合效果上来看,并没有显示出依赖性模型较之前的诸多模型有显著的改进(图 4.8)。在模型中追加点层次、二元组层次以及几何加权层次的统计项都会有助于模型拟合效果的提升。莫瑞斯和他的同事(Morris et al., 2008)针对有向网络可能的统计项和参量归纳了一个完整列表,希望进一步了解可选择统计项情况的研究人员可以参考这一资料。

表 4.4　波莫纳湖网络依赖性模型拟合优度测量结果

```
==========================
Summary of model fit
==========================

Formula:   lake ~ edges + nodeifactor("Location") + gwdsp(0.1, T) + gwesp(0.1,
    T)

Iterations:  20

Monte Carlo MLE Results:
                         Estimate Std. Error  MCMC %   Lower      OR   Upper  p-value
edges                    -1.78264   0.73643  1.00000 0.03971 0.16819  0.712  0.015965 *
nodeifactor.Location.NL  -0.78685   0.21719  0.00000 0.29744 0.45528  0.697  0.000331 ***
gwdsp.fixed.0.1          -0.26487   0.07223  0.00000 0.66602 0.76731  0.884  0.000281 ***
gwesp.fixed.0.1           1.64015   0.58406  1.00000 1.64115 5.15597 16.198  0.005242 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null  Deviance: 526.792  on 380  degrees of freedom
 Residual Deviance: 461.508  on 376  degrees of freedom
          Deviance:  65.284  on   4  degrees of freedom
AIC: 469.51    BIC: 485.27
```

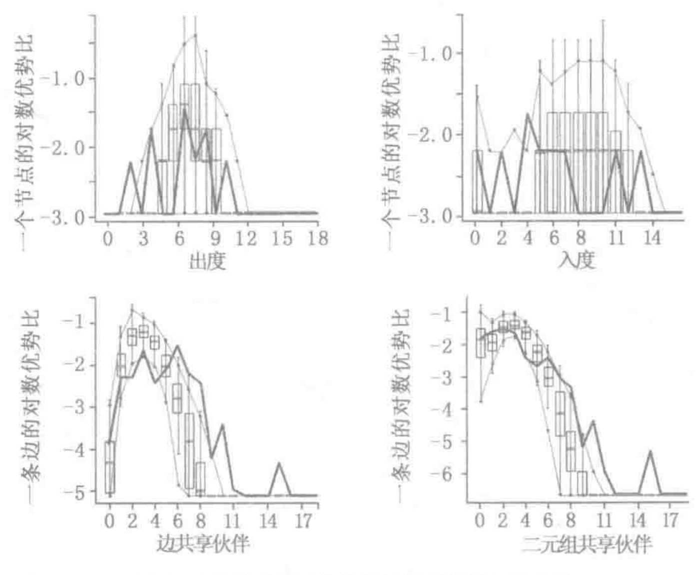


图 4.8 波莫纳湖网络依赖性模型的拟合优度图

与地方卫生机构网络一致，有向网络的模型也可以用来预测一个网络中链接关系形成的概率。然而，在进行预测概率的统计估计时，需要根据部分统计项的特征进一步区分链入关系和链出关系。

$$P(Y) = \text{logistic}\begin{pmatrix} -1.78 * \delta_{\text{edges}} - 0.79 * \delta_{\text{nonlocal-indegree}} + \\ 1.64 * \delta_{\text{GWESP}} - 0.26 * \delta_{\text{GWDSP}} \end{pmatrix}$$

$$\delta_{\text{GWESP, GWDSP}} = (1 - e^{-0.1})^{\eta_{\text{ESP, DSP}}} = 0.095^{\eta_{\text{ESP, DSP}}}$$

可以将上述结果代入到模型中进而预测某些情况下协作关系建立的概率。

案例 1：当 ESP 为 2 以及 DSP 为 4 时，预测从一个本地组织链入到一个非本地组织的概率。

$$P(Y) = \text{logistic}(-1.78 * 1 - 0.79 * 1 + 1.64 * 0.095^2 - 0.26 * 0.095^4)$$

$$P(Y) = \text{logistic}(-2.56)$$

$$P(Y) = 0.07$$

案例2：当ESP为2以及DSP为4时，预测从一个本地组织链入一个本地组织的概率。

$$P(Y) = \text{logistic}(-1.78 * 1 - 0.79 * 0 + 1.64 * 0.095^2 - 0.26 * 0.095^4)$$

$$P(Y) = \text{logistic}(-1.77)$$

$$P(Y) = 0.15$$

案例2中建立链接的概率为15%，这一概率较案例1中的7%要高出许多。上述结果显示的差异与此前在混合矩阵显示的差异是一致的，说明本地组织比非本地组织接收的链接多。

第 2 节 | 将二元组和网络协变量作为自变量

除了在地方卫生机构网络和波莫纳湖网络中介绍的点层次、二元组层次以及几何层次的自变量外，统计模型还可以采用其他的网络或二元组属性（例如，两节点之间的地理距离）作为网络模型的自变量。与此前所介绍的网络不同，当利用其他网络或者二元组属性作为模型的自变量时，模型中的自变量可以赋值，即关系所赋的值不限于0和1。

举例而言，R中包含的科尔曼（Coleman）交友数据集就包含了两个交友网络子集，这两个网络数据均来源于20世纪50年代后期，对伊利诺伊州一个小型高中的73个男孩之间交友关系进行的调研。该调研分别在第一年的秋季和第二年的春季对这73个男孩进行了调查。问题涉及"你在学校中最经常交往的同学是谁"。调查结果显示：在秋季调查时，朋友关系网呈现出两组男孩群体；而春季再调查时，朋友关系网络聚集为了一个单一的群体（Command 53；图4.9）。

注意，在春季交友关系网络统计摘要信息表中，包含了一个条目——边属性（edge attributes，参见表4.5阴影部

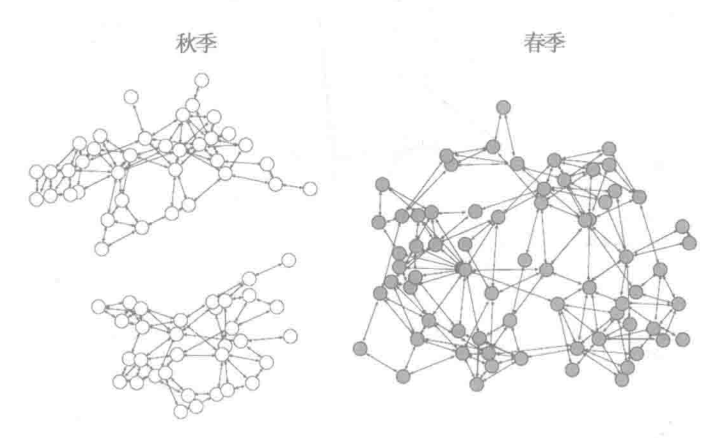


图 4.9　科尔曼的交友网络

表 4.5　春季时科尔曼交友网络的统计摘要表(包含对应的秋季边属性)

```
Network attributes:
  vertices = 73
  directed = TRUE
  hyper = FALSE
  loops = FALSE
  multiple = FALSE
  bipartite = FALSE
 total edges = 263
   missing edges = 0
   non-missing edges = 263
 density = 0.05003805

Vertex attributes:
  vertex.names:
   character valued attribute
   73 valid vertex names
Edge attributes:

 fall:
   numeric valued attribute
   attribute summary:
  Min.  1st Qu.  Median   Mean 3rd Qu.   Max.
0.0000  0.0000  1.0000  0.5323  1.0000  1.0000
```

续表

```
Network edgelist matrix:
      [,1][,2]
  [1,]  15   1
  [2,]  17   1
...
```

分)。本例中,摘要信息表显示了单一的边属性:秋季。该属性的含义是:在秋季时,如果两个男孩是朋友,则二元组对应的值为1;反之,则为0。

为了检验秋季的朋友关系是否可以用于预测春季的朋友关系,我们以秋季的朋友网络作为统计模型的预测变量来预测春季网络。一个仅包括边统计项的春季网络零模型,其AIC和BIC测量结果分别是2 090和2 096.5,当增加了一个秋季的朋友网络作为预测变量之后,模型的拟合效果出现了

表 4.6 利用秋季的科尔曼交友网络预测春季的朋友关系网络

```
==========================
Summary of model fit
==========================

Formula:  spring ~ edges + edgecov(fall)

Iterations:  20

Monte Carlo MLE Results:
              Estimate Std. Error MCMC %    Lower       OR  Upper p-value
edges         -3.68276    0.09129     NA  0.02103  0.02515   0.03 <1e-04 ***
edgecov.fall   3.98968    0.15870     NA 39.59179 54.03741  73.75 <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null  Deviance: 7286.4  on 5256  degrees of freedom
 Residual  Deviance: 1486.2  on 5254  degrees of freedom
           Deviance: 5800.1  on    2  degrees of freedom

AIC: 1490.2    BIC: 1503.4
```

极大的改进。结果显示：秋季的朋友关系对于春季朋友关系的形成具有显著且正向的影响（OR=54.0；95%CI=39.6－73.8）。所以，如果在秋季的科尔曼交友网络中的学生A已经显示与学生B建立了朋友关系，那么，学生A和学生B在春季时保持朋友关系的概率就是之前没有建立朋友关系的概率的54倍（Command 54；表4.6）。因此，可以说边属性可以与节点属性、几何加权统计项一样被纳入到模型中来。



第5章

结论与建议

本书首先是从一些网络实例开始的，如描述一个在线交友网络的幸福模式，以及一个由青年人构成的性关系网络中梅毒传播事件的事例。通过上述实例，我们认识到，在这些情景中，网络分析方法正在改变我们对于"关系是如何改变我们生活的各个方面"的理解。实际上，在早期的网络研究中，可视化方法以及描述性统计的方法都缺乏对网络中的关系模式之下所潜在的社会力量进行检验的功能。例如，在梅毒传播事件中青年人之间的性关系模式，就可以部分通过药物使用人员之间的同质性倾向特征来解释，于是，对应的干预策略就不仅要对药物使用行为进行干预，也同时要对青年之间的性行为进行干预。如果理解网络结构的目的在于增加或者减少网络中关系的形成，那么，洞悉网络结构之下所隐含的社会力量就是十分重要的环节。

在地方卫生机构模型中，本书对同质性假设的检验仅限于组织特征。然而，实际上，个体的行动也伴随着同质性特征，例如，抽烟行为和体育活动。许多采用标准网络方法进行的研究发现：在年轻人中，吸烟者更易于与吸烟者交朋友(Mercken, Snijders, Steglich & de Vries, 2009; Pearson, Steglich & Snijders, 2006)。因此，在保持其他结构特征不变

的情况下，网络中同质性特征对理解关系的形成提供了新的观察视角。正因为如此，本书已经将与行为相关的同质性假定正式纳入到指数随机图模型（ERGM）方法检验范畴中来。例如，最近德拉哈耶及罗宾斯等人（de la Haye, Robins, Mohr & Wilson, 2010）开展的一项研究就采用了指数随机图模型方法，该方法在青年人交友网络的基础上增加了一个结构效应，用于检验与肥胖疾病有关行为。德拉哈耶和他的同事发现：在控制了其他结构影响因素（如互惠性、流行性、扩展性、传递闭包以及多元连接性）之后，仍发现同质性特征的存在。这一发现有可能会改变公共健康管理者对于应如何防治青少年肥胖问题的认知，该研究建议应利用同辈压力以及其他社交策略的方法，引导肥胖青少年进行健康饮食和参与体育锻炼。

本书中描述的这些工具也为解决许多广泛存在的、持续且复杂的问题提供了新的思路。例如，由于在室内使用炉灶排放废气对人身体有害，因此，全球有30亿人口正在面临这种有害气体的疾病和死亡威胁，这些人主要是穷人（Yadama, Schechtman, Biswas, Castro & Chalise, 2013）。印度实施了一项旨在改善室内空气的干预措施，1985年到2000年间，给平民发放了数以百万计的低排放炉灶。然而，根据2002年的一项评估报告显示，85%的炉灶并没有得到使用，因此，这项干预措施被认为失败了。而通过对仅相隔10千米的两个村庄的对比研究发现，两个村庄在使用更清洁的炉灶方面具有惊人的差异。加图那（Gatuna）村显示有90%的居民采用了这种更清洁烹饪技术，而加里姆（Jalim）村则仅有10%的居民采用了这种更清洁的烹饪技术（Frandos, Fern, Yadama

& Bhatia, 2012)。对这两个村庄居民关系网络结构的思考,使我们认识到为什么在应用清洁炉灶使用上两个村庄会具有如此大的差异。例如,也许在加图那村采用清洁炉灶的居民之间比加里姆村采用清洁炉灶的居民之间具有更紧密的联系。如果真是这样,那么就应该建议采取一项策略,即应当在那些清洁炉灶采纳率低的村庄找到那些具有较广泛沟通关系的居民,然后,给这些居民以更加丰富的资源与技术支持,优先鼓励这些居民使用清洁炉灶。

统计学家和网络科学家尽管经过了几十年的努力,但本书中所描述的统计网络模型的发展和应用也才刚刚开始。正如斯蒂芬·霍金(Stephen Hawking)曾描述的那样,科学和社会正在步入一个"复杂性的世纪"。利用上述模型可以帮助我们理解与解释复杂性问题,这一点十分重要。记住,没有一个单一模型能够适合所有的网络研究,因此,无论是学生、应用科学家,还是其他开始采用统计网络模型进行研究的人们,均应该在其开展建模初期就考虑如下建议:

1. 在开始进行统计建模之前,先利用图形可视化和描述性统计方法对数据进行观察。通过直方图、混合矩阵以及网络可视化等方法对相关节点属性之间诸如同质性的模式进行观察。至少应观察网络中的度分布以及三角形数量特征对传递性进行评估,因为,观测网络的这两个特征经常与随机网络有较大区别。利用直方图和统计摘要表的方式对边共享伙伴和二元组共享伙伴分布进行评价,这一点可能有助于我们识别观测网络与随机网络在传递性模式上的差异。注意对传递性和其他网络结构特征的考察对于研究目的可能并不总是相关的,也不一定总是有效的。但正如我们在非

网络的研究中那样，研究人员在建模前应先利用探索性分析，逐步筛选出那些最适合研究问题的模型和方法。与结构性测量方法相比，针对度分布以及共享伙伴分布的测量（确保随机网络与所观测网络保持同等规模与密度）有助于我们识别随机网络与观测网络的差异，这一点可以帮助模型选择相应的变量。以上述结果为指导从而建立和精炼研究问题与假设，最终，指导模型的建立。

2. 建立一个零模型并且评价其拟合优度。

3. 有向模型中增加包括主效应和二元组层次属性特征等局部统计项，从而检验局部过程如何影响网络的结构。如果存在与主效应或者（二元组层次之间的）交互效应相关的假设，那么，首先应该增加主效应统计项，然后，增加交互效应统计项；通过观察上述模型中各个系数的显著性和方向，将前一阶段模型拟合的结果与期望值进行比较，然后再修正。同时，在整个分析过程中，不断观察统计及图形的拟合优度测量结果，并与零模型的结果进行比较。

4. 以前一阶段模型拟合时发现的观测网络与随机网络差异为基础，如有必要，可以向模型继续增加一些结构限制条件。如限定老化次数（burn-in）、样本规模以及本书中所提取的其他环境因素等。以 $\alpha=0.1$ 为初始条件，尝试采用多个 α 值，或者使用曲线指数族（CEF）策略来估计一个合适的 α 值。利用 MCMC 拟合诊断以及拟合图形测量来识别模型可能出现的近似退化和无法收敛问题，如觉得有必要则需要改进模型。检查模型统计项对应系数的显著性和方向性，确保这些统计项是符合逻辑的，并将该模型与之前已经建立的模型进行图形拟合优度比较，选择拟合度最高的模型、最符合

逻辑的模型作为解释并报告相应结论的模型。

5. 最终，帮助其他社会科学家建立对于网络以及统计网络模型的意识，如有必要则采用图形化的方式来解释模型建立的策略、模型的选择以及解释最终的模型。

对于那些旨在将指数随机图模型纳入到他们日常工具方法中的学者而言，可以考虑增加至少两个目前活跃的邮件列表(listservs)，以便能够跟上这个快速发展领域的新发展。首先，是国际社会网络分析学会(INSNA)的邮件列表，该学会是一个由网络分析方法人员构建的团体，对应的邮件列表(SOCNET)对于所有与网络有关的问题均开展了讨论，该邮件列表也是一个发现最新社会网络研究进展、各领域关键文献，以及了解各相关领域最活跃的研究人员的优秀信息来源。如何加入 SOCNET 邮件列表的说明可以在 INSNA 的网站上获取(http://www.insna.org/)。其次，statnet 邮件列表也是极为活跃的，statnet 的开发人员通常会很快地对 statnet 用户的提问作出反馈。如何加入 statnet 邮件列表的说明可以在 statnet 的网站上获取(http://statnet.org/)。

进一步阅读和资源

与 statnet 相关的资料。有大量关于 R-statnet 资源可供学者们进一步阅读，这些资料大多数都可以在 statnet 的网站上获取(http://statnet.org/)。2008 年 5 月的《统计软件杂志》专刊包括了对于理解和使用 statnet 极为有用的文章。这期专刊中的所有文章都可以通过网络公开获取(http://www.jstatsoft.org/v24/)。最终，statnet 的开发人员也经常

会参加一些网络会议，如国际社会网络分析大会以及密歇根大学高校校际政治与社会研究联盟（ICPSR）夏季项目（http://www.icpsr.umich.edu/）。同时，前述的邮件列表中也会列出一些其他的培训机会。

关于 ergm 的相关资料。2008 年 5 月的《统计软件杂志》专刊主要是对于 statnet 中指数随机图模型的评价，同时也包含对近期指数随机图模型发展的思考。2007 年 5 月的《社会网络》期刊中有一期针对指数随机图模型的专刊（Volume 29，Issue 2），该期包含了大量有用的文章，许多文章都可以在本书的参考文献部分找到。2011 年出版的《SAGE 社会网络分析手册》也包含一些对于指数随机图模型和其他统计网络模型有用的章节。



附　录

在有向网络中存在 16 种可能的三元组构成,每一种三元组构成都对应了一项标识符。图 A.1 显示了这 16 种三元组构成及其对应的标识符。

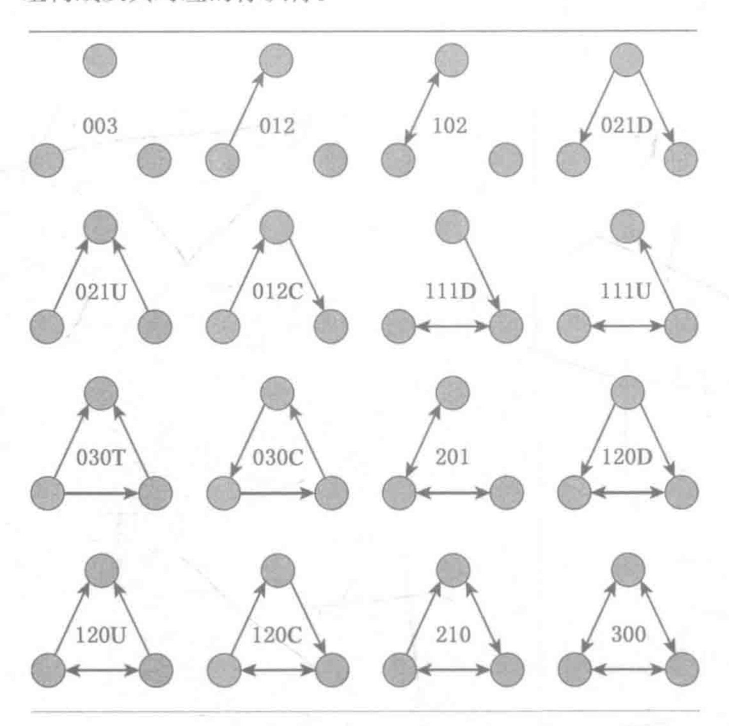


图 A.1 有向网络的三元组类型

参考文献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akadémiai Kiado.

An, W. (2011). Models and methods to identify peer effects. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 514–532). London, UK: Sage.

Anderson, C. J., Wasserman, S., & Faust, K. (1992). Building stochastic blockmodels. *Social Networks, 14,* 137–161.

Barnes, J. A. (1972). Social networks. *Addison-Wesley Module in Anthropology, 26,* 1–29.

Beatty, K. B., Harris, J. K., & Barnes, P. (2010). The role of inter-organizational partnerships in health services provision among rural, metropolitan, and urban local health departments. *Journal of Rural Health, 26,* 248–258.

Berkman, L. F., & Syme, S. L. (1979). Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda county residents. *American Journal of Epidemiology, 109*(2), 186–204.

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., & Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science, 3*(5), 388–397.

Bollen, J., Goncalves, B., Ruan, G., & Mao, H. (2011). Happiness is assortative in online social networks. *Artificial Life, 17*(3), 237–251.

Box, G. E. P., & Draper, N. R. (2007). *Response surfaces, mixtures, and ridge analyses* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Buchanan, M. (2002). *Nexus: Small worlds and the groundbreaking science of networks*. New York: W.W. Norton.

Burt, R. S. (1987). A note on strangers, friends, and happiness. *Social Networks, 9,* 311–331.

Butts, C. T. (2008). Network: A package for managing relational data in R. *Journal of Statistical Software, 24*(8), 1–36.

Caulkins, D. (1981). The Norwegian connection: Eilert Sundt and the idea of social networks in 19th century ethnology. *Connections, 4*(2), 28–31.

Centers for Disease Control and Prevention. (1998). Outbreak of primary and secondary syphilis—Guilford county, North Carolina, 1966–1997. *Morbidity and Mortality Weekly Report, 47*(49), 1070–1073.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*(434), 883–904.

Cranmer, S. J., & Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Political Analysis, 19*(1), 66–86.

de la Haye, K., Robins, G., Mohr, P., & Wilson, C. (2010). Obesity-related behaviors in adolescent friendship networks. *Social Networks, 32*(3), 161–167.

Ennett, S. T., & Bauman, K. E. (1993). Peer group structure and adolescent cigarette smoking: A social network analysis. *Journal of Health and Social Behavior, 34*(3), 226–236.

Erdős, P., & Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen), 6,* 290–297.

Field, A. P. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage.

Frandos, A., Fern, S., Yadama, G., & Bhatia, V. (2011, July). *Uptake of alternative energy technology by energy poor households in rural Rajasthan, India*. Paper presented at the International System Dynamics Conference, Washington, DC.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association, 81*(395), 832–842.

Freeman, L. C. (1996). Some antecedents of social network analysis. *Connections, 19*(1), 39–42.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, BC: Empirical Press.

Freeman, L. C. (2011). The development of social network analysis—with an emphasis on recent events. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 26–41). London, UK: Sage.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., & Morris, M. (2008). A statnet tutorial. *Journal of Statistical Software, 24*(9), 1–27.

Goodreau, S. M., Kitts, J. A., & Morris, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography, 46*(1), 103–125.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*, 201–233.

Green, L. W. (2006). Public health asks of systems science: To advance our evidence-based practice, can you help us get more practice-based evidence? *American Journal of Public Health, 96*(3), 406–409.

Hall, J. A., & Valente, T. W. (2007). Adolescent smoking networks: The effects of influence and selection on future smoking. *Addictive Behaviors, 32*(12), 3054–3059.

Harris, J. K., Baker, E. A., Barnidge, E., McGee, L., Motton, F., Rose, R., Roche, J., et al. (2012, March). *Employment networks in a high-unemployment rural area*. Paper presented at the International Network for Social Network Analysis Sunbelt Conference, Redondo Beach, CA.

Harris, J. K., Carothers, B. J., Wald, L. M., Shelton, S. C., & Leischow, S. J. (2012). Interpersonal influence among public health leaders in the United States Department of Health and Human Services. *Journal of Public Health Research, 1*(1), 67–74.

Harris, J. K., Cyr, J., Carothers, B. J., Mueller, N. B., Anwuri, V. V., & James, A. I. (2011). Referrals among cancer services organizations in an underserved urban area. *American Journal of Public Health, 101*(7), 1248–1252.

Harris, J. K., Luke, D. A., Burke, R. C., & Mueller, N. B. (2008). Seeing the forest and the trees: Using network analysis to develop an organizational blueprint of state tobacco control systems. *Social Science & Medicine, 67*(11), 1669–1678.

Harris, J. K., Luke, D. A., Zuckerman, R. B., & Shelton, S. C. (2009). Forty years of secondhand smoke research: The gap between discovery and delivery. *American Journal of Preventive Medicine, 36*(6), 538–548.

Hirsch, G. B., Levine, R., & Miller, R. L. (2007). Using system dynamics modeling to understand the impact of social change initiatives. *American Journal of Community Psychology, 39*(3–4), 239–253.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association, 76*(373), 33–50.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks, 29*(2), 216–230.

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association, 103*(481), 248–258.

Hunter, D. R., & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics, 15*(3), 565–583.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). Ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software, 24*(3), 1–29.

Karonski, M. (1982). A review of random graphs. *Journal of Graph Theory, 6*(4), 349–389.

Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 141–182). New York: Academic Press.

Krackhardt, D. (1987). QAP partialling as a test of spuriousness. *Social Networks, 9*(2), 171–186.

Krebs, V. (2000). Working in the connected world: Book network. *International Association for Human Resource Information Management, 4*(1), 87–90.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics, 6,* 1100–1128.

Krivitsky, P. N., & Handcock, M. S. (2008). Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software, 24*(5), 1–23.

Leischow, S. J., Best, A., Trochim, W. M., Clark, P. I., Gallagher, R. S., Marcus, S. E., et al. (2008). Systems thinking to improve the public's health. *American Journal of Preventive Medicine, 35*(Suppl. 2), S196–S203.

Luke, D. A. (2005). Getting the big picture in community science: Methods that capture context. *American Journal of Community Psychology, 35*(3–4), 185–200.

Luke, D. A., & Harris, J. K. (2007). Network analysis in public health: History, methods, and applications. *Annual Review of Public Health, 28*, 69–93.

Luke, D. A., Harris, J. K., Shelton, S., Allen, P., Carothers, B. J., & Mueller, N. B. (2010). Systems analysis of collaboration in 5 national tobacco control networks. *American Journal of Public Health, 100*(7), 1290–1297.

Luke, D. A., & Stamatakis, K. A. (2012). Systems science methods in public health: Dynamics, networks, and agents. *Annual Review of Public Health, 33*, 357–376.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27,* 415–444.

Mercken, L., Snijders, T. A. B., Steglich, C., & de Vries, H. (2009). Dynamics of adolescent friendship networks and smoking behavior: Social network analyses in six European countries. *Social Science & Medicine, 69*(10), 1506–1514.

Moreno, J. (1953). *Who shall survive? Foundations of sociometry, group psychotherapy and sociodrama* (2nd ed.). Beacon, NY: Beacon House. (Original work published 1934)

Morris, M., Handcock, M. S., & Hunter, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software, 24*(4), 1–24.

Myers, D. G., & Diener E. (1995). Who is happy? *Psychological Science, 6*(1), 10–19.

Pattison, P., & Robins, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology, 32*(1), 301–337.

Pearson, M., Steglich, C., & Snijders, T. A. B. (2006). Homophily and assimilation among sport-active adolescent substance users. *Connections, 27*(1), 47–63.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11.

Rivera, M. T., Soderstrom, S. B., & Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology, 36,* 91–115.

Robins, G. (2011). Exponential random graph models for social networks. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 484–500). London, UK: Sage.

Robins, G. L., Snijders, T. A. B., Wang, P., Handcock, M. S., & Pattison, P. E. (2007). Recent developments in exponential random graph (p*) models for social networks. *Social Networks, 29*, 192–215.

Rothenberg, R. B., Sterk, C., Toomey, K. E., Potterat, J. J., Johnson, D., Schrader, M., et al. (1998). Using social network and ethnographic tools to evaluate syphilis transmission. *Sexually Transmitted Diseases, 25*(3), 154–160.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Scott, J., & Carrington, P. J. (Eds.). (2011). *The SAGE handbook of social network analysis*. London, UK: Sage.

Seeman, T. E., Kaplan, G. A., Knudsen, L., Cohen, R., & Guralnik, J. (1987). Social network ties and mortality among the elderly in the Alameda county study. *American Journal of Epidemiology, 126*(4), 714–723.

Shumate, M., & Palazzolo, E. T. (2010). Exponential random graph (p*) models as a method for social network analysis in communication research. *Communication Methods and Measures, 4*(4), 341–371.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure, 3*(2).

Snijders, T. A. B. (2011a). Statistical models for social networks. *Annual Review of Sociology, 37*, 131–153.

Snijders, T. A. B. (2011b). Network dynamics. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 501–513). London, UK: Sage.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology, 36*(1), 99–153.

Snijders, T. A. B., van de Bunt, G. G., & Steglich, C. E. G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks, 32*, 44–60.

Ura, K., Alkire, S., Zangmo, T., & Wangdi, K. (2012). *A short guide to gross national happiness index*. Thimphu: The Centre for Bhutan Studies.

Valente, T. W. (2010). *Social networks and health: Models, methods, and applications*. New York, NY: Oxford University Press.

Valente, T. W., & Saba, W. P. (1998). Mass media and interpersonal influence in a reproductive health communication campaign in Bolivia. *Communication Research, 25*(1), 96–124.

Valente, T. W., & Vlahov, D. (2001). Selective risk taking among needle exchange participants: Implications for supplemental interventions. *American Journal of Public Health, 91*(3), 406–411.

van Duijn, M. A. J., & Huisman, M. (2011). Statistical models for ties and actors. In J. Scott & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 459–483). London, UK: Sage.

van Duijn, M. A. J., Snijders, T. A. B., & Zijlstra, B. J. H. (2004). p_2: A random effects model with covariates for directed graphs. *Statistica Neerlandica, 58*(2), 234–254.

Voorhees, C. C., Murray, D., Welk, G., Bimbaum, A., Ribisl, K. M., Johnson, C. C., et al. (2005). The role of peer social network factors and physical activity in adolescent girls. *American Journal of Health Behavior, 29*(2), 183–190.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika, 61*(3), 401–425.

Wasserman, S., & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 148–161). New York, NY: Cambridge University Press.

Yadama, G., Schechtman, K. B., Biswas, P., Castro, M., & Chalise, N. (2013). Indoor air pollution and respiratory health: A transdisciplinary vision. In D. Haire-Joshu, & T. McBride (Eds.), *Transdisciplinary public health: Research, methods, education and practice* (pp. 319–340). San Francisco, CA: Jossey-Bass.

译名对照表

actor-focused branch	以行动者为中心的研究分支
Akaike information criterion(AIC)	赤池信息准则
antipreferential attachment	异配倾向
assumptions, general linear models and	广义线性模型的假设
attractiveness	吸引性
Bayesian information criterion(BIC)	贝叶斯信息准则
behaviors, relationships and	关系与行为
binary networks	二值网络
Center for Disease Control and Prevention	美国疾病控制与预防中心
change statistic	变化统计
Coleman friendship networks	科尔曼交友网络
confidence intervals(CIs)	置信区间
constraints, refining model using	利用约束条件对模型进行精炼
correlation coefficient	相关系数
curved exponential family(CEF)	曲线指数族(CEF)
data	数据
accessing in ERGM software	在 ERGM 软件包中获取数据
egocentric network	自我中心网数据
epidemiologic	流行病学数据
exploring network(ERGM)	利用 ERGM 对网络数据进行探索
degeneracy, in network modeling	网络建模中的近似退化
degree distributions	中心度分布
deviance	离差
differential attractiveness	差异化吸引性
directed networks	有向网络
distibution, of triangles per network	网络的三角形数量分布
dyadic dependence models	二元依赖性模型
dyadic independence models	二元独立性模型

dyadic network covariates, as predictors	将二元网络协变量作为自变量
dyads, in networks	网络中的二元组
dyad types	二元组类型
dyadwise shared partners(DSP)	二元组共享伙伴
edge attributes	边属性
edges term	边统计项
edgewise shared partners(ESP)	边共享伙伴
egocentric network data	自我中心网络数据
epidemiologic data	流行病学数据
exponential decline	指数递减
exponential family model, curved.	曲线指数族模型
exponential random graph modeling(ERGM)	指数随机图模型
exponential random graph modeling(ERGM) development	指数随机图模型的构建
accessing data	获取数据
adding node attributs	增添节点属性
constraints, refining model	模型精炼的限定条件
curved exponential family model	曲线指数族模型
dependence model, interpreting results	依赖关系模型的解释结果
dependence terms	依赖关系统计项
interaction terms	交互统计项
MCMC model diagnostics	马尔科夫链模型诊断
model fit	模型拟合
model selection	模型选择
network data, exploring	对网络数据进行探索性分析
null model	零模型
obtaining/preparing software	获取/准备软件
probabilities, predicting	预测概率
frequency, network	网络的频次
friendships, happiness and	幸福与友谊

geometrically weighted degree distribution(GWD)	几何加权度分布
geometrically weighted dyadwise shared partners(GWDSP)	几何加权二元组共享伙伴
geometrically weighted edgewise shared partners(GWESP)	几何加权边共享伙伴
GNH index, see Gross National Happiness(GNH) index	国民幸福总值指数
goodness-of-fit	拟合优度
graphic diagnostics	(对拟合优度采取的)可视化诊断
graphic examination of degree	中心度的可视化验证
Gross National Happiness(GNH) index	国民幸福总值指数
happiness, social networks and	社会网络与幸福
high-degree nodes	高中心度节点
higher-order dependence models	高序依赖性模型
homophily, in models	模型中的同质性
individuals, networks and	网络与个体
interaction terms, adding	增加交互统计项
International Network for Social Network Analysis(INSNA)	国际社会网络分析学会
Inter-university Consortium for Political and Social Research(ICPSR)	高校校际政治与社会研究联盟
in-ties	链入关系
Journal of Statistical Software	《统计软件杂志》
keep argument	"keep"参数
linear incline	线性倾斜
listserv, statnet	statnet 的邮件列表服务
log-likelihood(LL)	对数似然比
log odds	对数优势比
main effects model, predicting probabilities and	预测概率与主效应模型

Markov chain Monte Carlo(MCMC)	马尔科夫链蒙特卡洛方法
Markov dependence assumption	马尔科夫依赖性假定
matrice, mixed	混合矩阵
min column	最小值列
mixed matrices	混合矩阵
model development, recommendations for	模型构建建议步骤
model effects	模型效应
model fit, ERGM and	ERGM 和模型拟合
model selection, ERGM and	ERGM 和模型选择
multilevel modeling	多层次建模
multilevel regression model	多层次回归模型
multinet software multinet	multinet 软件
multiple regression quadratic assignment procedure	多元回归二次指派程序
mutual term	交互统计项
National Association of County and City Health Officials(NACCHO)	美国国家城镇卫生官员协会
network covariates, as predictors	网络协变量作为自变量
network density	网络密度
network lexicon	网络术语
networks	网络
makeup of,	网络的构成
observed, GWD and	几何加权二元组和观测网络
online friendships and happiness	在线交友和幸福感网络
organizations, connections and	组织、联系与网络
relationship patterns and	关系类型及网络
network size	网络规模
network statistic	网络统计
network structures	网络结构
network tools, ERGM	ERGM 网络工具
nodcov main effect	nodcov 主效应

node attributes, adding	增加节点属性统计项
nodefactor	节点因素
nodes, size/shape	节点规模/形状
nonuniform	非均匀
nonuniform degree distribution	非均匀的度分布
null hypothesis	零假设
null model	零模型
obs column	观测值列
odds ratios(ORs)	优势比
online friendships, happiness and	幸福感与在线交友网络
organizations, connections, networks and	网络、组织与联系
partial conditional dependence	部分条件依赖
p_1 model	p_1 模型
Pnet software	Pnet 软件
predictors	自变量
dyadic and network covariates as	将二元组以及网络协变量作为自变量
values of	自变量值
probabilities	概率
predicting for CEF models	预测 CEF 模型的概率
predicting with ERGM	利用 ERGM 预测概率
proportion, nodes	节点比例
PSPAR software	PSPAR 软件
quadratic assignment procedure(QAP)	二次指派程序
random network	随机网络
receiver terms	接收者效应统计项
reciprocity	互惠性
relationship patterns, sexual contact	性接触关系模式
R Project for Statistical Computing(website)	面向统计计算的 R 项目(网页)
RSIena software	RSIena 软件

R-statnet, software	R-statnet 软件包
SAGE Handbook of Social Network Analysis	SAGE 社会网络分析手册
sender terms	发送者效应统计项
sexually transmitted disease, relationship patterns and	关系模式和性传播疾病
simple random graphs	简单随机图
simulated networks	仿真网络
social circuit dependence	社交圈依赖关系
social network analysis(SNA)	社会网络分析(SNA)
Social Networks(journal)	社会网络(期刊)
social relations model(SRM)	社会关系模型(SRM)
sociogram	社群图
SOCNET, INSNA listserv	SOCNET, INSNA 的邮件列表服务
spatial statistics	空间统计
statistical network models	统计网络模型
dyadic dependence models	二元依赖性模型
dyadic independence models	二元独立性模型
ERGM, development of	ERGM 的发展
higher-order dependence models	高序依赖性模型
simple random graphs	简单随机图模型
statnet(website)	statnet 网站
stochastic block models	随机块模型
subjective well-being(SWB)	主观幸福感
terms, p_1 model	p_1 模型统计项
tie-focused branch	以关系为中心的研究分支
transivity	传递性
triad types	三元组类型
triangles per network	网络的三角形数
value column	分值列
values of predictors	自变量值

vertex attribute names	顶点的属性名称
vertex size/shape	顶点的规模/形状
Wald test	沃尔德检验
weighting parameter	加权参数
Who Shall Survive? (Moreno)	《谁将生存?》(莫雷诺)

译后记

本书的译稿是国家自然科学基金项目青年项目“基于指数随机图模型的专利引用关系形成影响因素及机理研究”的阶段性研究进展，项目编号：71403256。

另外，翻译过程中，原书作者詹宁·哈瑞斯（Jenine Harris）给予了我巨大的鼓励与帮助，耐心地解释了我提出的问题，并慷慨地提供了书中原始的附图资料。在此，我对哈瑞斯女士致以最真挚的感谢。

An Introduction to Exponential Random Graph Modeling

上海市版权局著作权合同登记号:图字 09-2014-688

格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic回归入门
39. 解释概率模型：Logit、Probit以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic回归中的交互效应
46. 应用回归导论
47. 档案数据处理：研究“人生”
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论